Détection d'homologies lointaines à faibles identités de séquences : Application aux protéines de la signalisation des dommages de l'ADN

Vincent Meyer

Université Paris VII – Denis Diderot

Ecole Doctorale Inter///BIO

Thèse présentée pour obtenir

Le GRADE de DOCTEUR en SCIENCES

par

Vincent MEYER

Sujet :

Détection d'homologies lointaines aux faibles identités de séquences : Application aux protéines de la signalisation des dommages de l'ADN.

Thèse dirigée par Bernard GILQUIN,

Soutenue le 26 janvier 2007 devant la commission d'examen :

Madame Catherine ETCHEBEST (présidente de jury)

Monsieur Olivier POCH (rapporteur)

Monsieur Gilles LABESSE (rapporteur)

Madame Isabelle CALLEBAUT (examinatrice)

Monsieur Bernard GILQUIN (directeur de thèse)

TABLE DES MATIERES

ABREVIATIONS

A

ADN: Acide DésoxyriboNucléique
ATM: Ataxia-Telangiectasia Mutated

B

BAliBASE: Benchmark Alignement data BASE
BLAST: Basic Local Alignment Search Tool
BLOSUM: BLOcks SUbstitution Matrix
BRCT: Breast Cancer Carboxy-Terminal Protein

C

CATH: Class, Architecture, Topology and Homologous superfamily
CASP: Critical Assessment of Techniques for Protein Structure Prediction
CAFASP: Critical Assessment of Fully Automated Structure Prediction
COMPASS: Comparison of Multiple Protein Alignements with Assessment of Statistical Significance
COACH: Comparison Of Alignement by Constructing Hidden markov model

D

DALI: Distance matrix ALIgnement
DDBJ: Dna Data Bank of Japan
DSSP: Database of structurally similar protein

E

EBI: European Bioinformatics Institute
EMBL: European Moleculary Biology Laboratory
EMBO: European Moleculary Biology Organisation

F

FHA: Forkhead Associated Domain
FSSP: Families of structurally similar protein

H

HCA: Hydrophobic Cluster Analysis
HMM: Hidden Markov Model
HHnoss: Utilisation du logiciel HHsearch dans l'assistance des prédictions de structure secondaire.

HOMSTRAD: HOMologous STRucture Alignement Database
HSP: High-scoring Segment Pairs
HHss: Utilisation du logiciel HHsearch avec l'assistance des prédictions de structure secondaire.
HSSM: Hidden semi-Markov Model

J
JIPID: Japan International Protein Information Database

L
LIF1: Ligase Interacting Factor 1

M
MAFFT: Multiple Alignement Based on Fast Fourier Transform
MIPS Martinsried Institute for Protein Sequences
Mre: Meiotic Recombination
MSP: pour Maximal-scoring Segment Pair
MUMMALS: Multiple sequence alignement improved by using hidden Markov models with local structural information.
MUSCLE: Multiple Sequence Comparison by log-expectation

N

NADP(H) : Nicotinamide adénine dinucléotide phosphate.
NBRF: National Biomedical Research Foundation.
NBS1: Nijmegen Breakage Syndrome 1
NCBI: National Center for Biotechnology Information
NEJ1: Non-homologous End Joining 1
NHEJ: Non-Homologous End Joining
NIH: National Institute of Health

P

PAM: Percentage of Acceptable points Mutation
PDB: Protein Data Bank
PFAM: Protein Family
ProbCons: Probabilistic Consitency-based Multiple alignement
PSI-BLAST : Position Spécific itérated BLAST
PSI-PRED : Position Spécific itérated PREDiction
PSSM : Position-Specific Score Matrix

Q

QC-COMP : Quasi-consensus comparison
Qdev: Qdeveloper
Qmod: Qmodeler

R
RAD"n": RADiation sensitive protein "n"
RCSB: Research Collaboratory for Structural Bioinformatics
RMN: Résonance Magnétique Nucléaire
ROC: Receiver Operating Characteristic

S
SMART : Simple Modular Architecture Research Tool
SABmark: Sequence Alignement Benchmark
SCOP Structural Classification of Proteins
SGD: Saccharomyces Genome Database
SIB: Swiss Institute of Bioinformatics
SMN: Survival MotoNeuron
SPEM: Sequence and secondary-structure Profiles Enhanced Multiple alignement)
SVM: Support Vector Machines

T
TrEMBL: Translated EMBL nucleotide sequence library

X
XRS2: X Ray Sensitive protein 2

Remerciements:

Je remercie tout d'abord l'ensemble des membres du jury d'avoir accepté d'évaluer l'ensemble de ces travaux.

Madame Catherine ETCHEBEST pour avoir accepté de présider ce jury.

Monsieur Olivier POCH et Monsieur Gilles LABESSE pour avoir évalué ce mémoire en qualité de rapporteurs.

Madame Isabelle CALLEBAUT pour avoir évalué ce mémoire en qualité d'examinatrice.

Je tiens aussi à remercier les personnes et organisations suivantes qui ont contribué à l'aboutissement de ces travaux.

Monsieur André Ménez: pour m'avoir accueilli au sein de son laboratoire et accepté d'assurer mon financement au cours cette dernière année.

L'Association Francaise contre les Myopathies: pour avoir financé mes 2 premières années de doctorat.

Bernard Gilquin: pour avoir accepté d'être mon directeur de thèse, m'avoir encadré et guidé au cours de la rédaction de ce mémoire.

Jean-Michel Neumann: pour m'avoir accueilli dans les locaux du SBGM.

Raphaël Guerois: pour m'avoir encadré au cours de ces 3 années et guidé au cours de ces travaux ainsi que lors de la rédaction de ce mémoire.

Sophie Zinn-Justin: pour avoir participé à mon encadrement et aux corrections de ce mémoire.

Hocine Madaoui, Thien An Nguyen, Emmanuelle Becker pour ces 3 années passées avec vous...les discussions scientifiques...et tout le reste ;)

Ma compagne Peggy Regulus, pour avoir supporté les distances et les emplois du temps qui nous ont séparés pendant ces 3 années.

Mes parents pour leur soutien au quotidien.

Et enfin à toutes les personnes non citées ici, qui d'une manière ou d'une autre, m'ont accompagné pendant ces 3 années....a vous aussi ...merci!

Chapitre I :Introduction

La bioinformatique est un domaine de recherche qui existe depuis plusieurs dizaines d'années. Récemment, l'augmentation vertigineuse du nombre de données générées en biologie associée au développement d'outils informatiques spécifiques à la recherche en biologie a peu à peu mis en lumière cette nouvelle discipline. La bioinformatique comprend un ensemble de concepts et de techniques nécessaires à l'interprétation de l'information génétique (séquences et expressions) et structurale (repliement 3-D, interactions). C'est le décryptage de la 'bio-information' ('Computational Biology' en anglais). La bioinformatique est donc une branche théorique de la biologie. Son but est d'effectuer la synthèse des données disponibles (à l'aide de modèles et de théories), d'énoncer des hypothèses généralisatrices (ex. : comment les protéines se replient), et de formuler des prédictions (ex. : prédire la fonction d'un gène).

La bioinformatique s'appuie sur des domaines de l'informatique « traditionnelle », que ce soit pour l'acquisition des données (instrumentation, robotique), leur archivage (bases de données) ou leur consultation (interface utilisateurs, Internet). Ces domaines ne sont pas spécifiques à la biologie (Figure 1), mais leur évolution rapide permet d'améliorer constamment l'analyse des données biologiques qui sont aujourd'hui générées de manière massive.

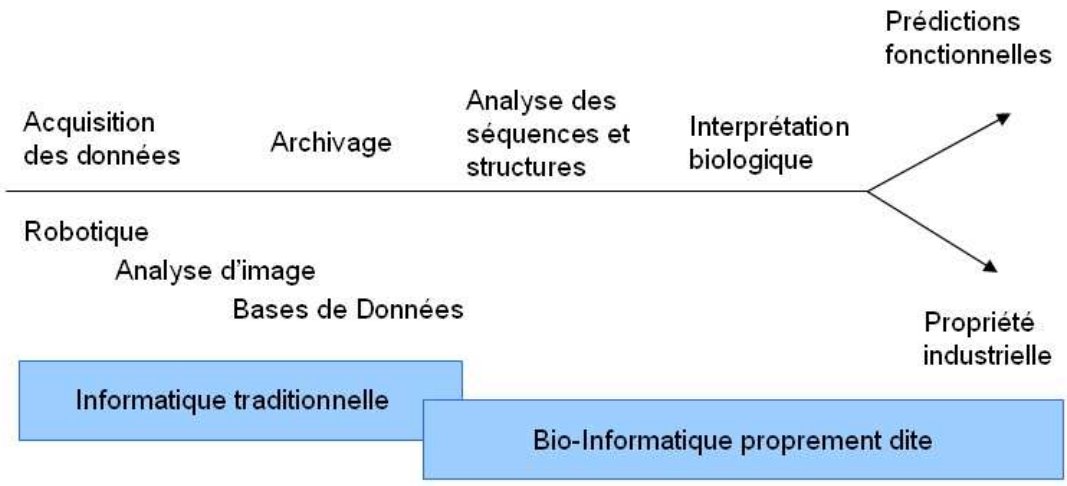


Figure 1 : D'après la figure extraite de « La bioinformatique: une discipline stratégique pour l'analyse et la valorisation des génômes »,JM. Claverie, S. Audic & C. Abergel, http://igs-server.cnrs-ms.fr .

La bioinformatique représente aujourd'hui un domaine en pleine expansion aussi bien dans le monde de la recherche que dans l'industrie. On retrouve ainsi plusieurs journaux dédiés à la bioinformatique, et des résultats apportés par la bioinformatique contribuent à l'argumentaire des demandes de brevets (ex : 'ce gène partage tel motif avec tel autre, a donc telle fonction probable, et peut donc être à la base de telle application pharmacologique').

Dans cette thèse, nous abordons l'analyse bioinformatique des séquences biologiques. Une telle analyse comporte en général trois aspects :

- **Compilation et organisation des données**. Cet aspect concerne essentiellement la création de bases de données. Certaines bases ont pour vocation de réunir le plus d'informations possible sans expertise particulière de l'information déposée alors que d'autres ciblent un domaine spécifique. Ces dernières bases sont généralement construites autour de thèmes et de fonctions précises comme l'ensemble des séquences d'une même espèce ou les séquences correspondant à une même fonction biologique. Elles sont constituées en amont de tout développement de méthodes d'analyse ou de prédiction.
- **Traitement systématique des séquences.** L'objectif principal est de repérer, parmi un grand nombre de séquences, une caractéristique structurale ou fonctionnelle intéressante. Ces traitements permettent, par exemple, l'identification de phases codantes sur une molécule d'ADN ou la recherche de similitudes entre une séquence et l'ensemble des séquences d'une base de données.
- **Elaboration de stratégies**. Le but est d'extraire des données biologiques, par une approche originale, de nouvelles informations qui seront ensuite intégrées dans des traitements standards. On peut donner comme exemples la mise au point de nouvelles matrices de substitution des acides aminés, la détermination de l'angle de courbure d'un segment d'ADN en fonction de sa séquence primaire, ou encore la détermination de critères spécifiques dans la définition de séquences régulatrices.

Notre objectif est de mettre au point un traitement bioinformatique permettant de détecter des similitudes structurales et éventuellement fonctionnelles entre des séquences *a*

priori très divergentes entre elles. Une telle approche permettra l'identification de domaines structuraux au sein de séquences ne ressemblant apparemment à aucune autre séquence connue. Elle facilitera ainsi l'annotation, la détermination de la structure 3D et l'analyse fonctionnelle de régions de ces séquences dites 'orphelines'. Je présenterai tout d'abord les banques de séquences et de structures 3D protéiques, ainsi que les méthodes permettant la comparaison et la classification des éléments de ces banques. Puis je décrirai les grandes lignes de la méthode que nous avons employée afin de mettre en évidence la présence de séquences structuralement proches entre des séquences dont la similarité n'a pas été caractérisée par un traitement classique.

I.1. Etat des lieux des séquences et structures connues

I.1.1. Les banques de séquences biologiques

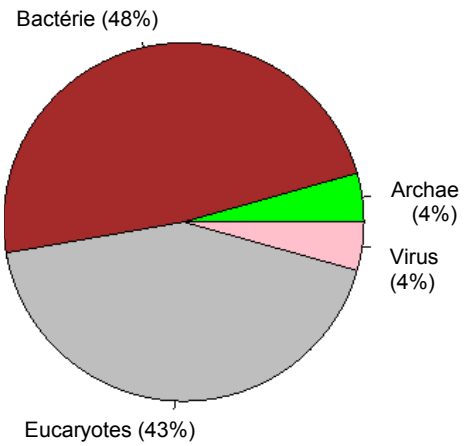
C'est au début des années 80 que les premières banques de séquences sont apparues sur l'initiative de plusieurs équipes. Très rapidement, avec l'augmentation de l'efficacité du séquençage, la collecte et la gestion des données ont nécessité une organisation plus conséquente. Plusieurs organismes ont pris en charge la production de telles bases de données. En Europe, une équipe financée par l'EMBO (European Moleculary Biology Organisation) s'est constituée pour développer une banque de séquences nucléiques (**EMBL** data library) et en assurer la diffusion . Cette équipe travaille au sein du Laboratoire Européen de Biologie Moléculaire qui, longtemps resté à Heidelberg, se trouve actuellement près de Cambridge au sein de l'**EBI** (European Bioinformatics Institute). Du côté américain, une banque de séquences nucléiques nommée **GenBank** et financée par le NIH (National Institute of Health) a été créée à Los Alamos . Cette base de données, au départ distribuée par la société IntelliGenetics, est maintenant diffusée par le **NCBI** (National Center for Biotechnology Information). Une collaboration entre ces banques européenne et américaine a été initiée relativement tôt, et s'est étendue en 1987 avec la participation de la DDBJ (Dna Data Bank of Japan) du Japon, pour donner naissance finalement en 1990 à un format unique de description des séquences dans les banques de données nucléiques (**The DDBJ/EMBL/GenBank feature table : Definitions, 1999**, http://www.ebi.ac.uk/embl/Documentation/FT_definitions/feature_table.html).

Parallèlement, deux banques principales de séquences protéiques ont été créées. La première, sous l'influence du National Biomedical Research Foundation (NBRF) à Washington, a produit une association de données issues du MIPS (Martinsried Institute for

Protein Sequences), de la base japonaise JIPID (Japan International Protein Information Database) et des données propres de la NBRF. Elle se nomme la Protein Identification Ressource (**PIR-NBRF**) . La deuxième, gérée par l'EBI et le SIB (Swiss Institute of Bioinformatics), a été constituée à l'Université de Genève à partir de 1986. Elle comprend une banque de données soigneusement annotée, la **Swissprot** , et une banque plus large annotée automatiquement, la **TrEMBL** (Translated EMBL nucleotide sequence library) . Aujourd'hui l'EBI, le SIB et le PIR ont uni leurs forces pour créer la base de données **UNIPROT** . La version non redondante de cette base de données, UniRef100, contient environ 3 millions et demi de séquences (Tableau 1). Si l'on analyse un sous-extrait soigneusement annoté de cette banque, UniProtKB Swiss-Prot, on constate qu'environ la moitié des séquences sont d'origine bactérienne (Figure 2). Ceci reste vrai pour l'ensemble des bases de données d'UNIPROT, et nous verrons qu'il faut en tenir compte lors de l'analyse des résultats d'une recherche sur cette banque.

UniProt	Release	8.0
Database		Entries
UniProtKB		3,170,612
UniProtKB/Swiss-Prot section:		222,289
UniProtKB/TrEMBL section :		2,948,323
UniRef100		3,511,676
UniRef90		2,254,474
UniRef50		1,148,123
UniParc		7,116,519

Tableau 1: Statistiques de la base de données UniProt : UniprotKB comprend l'ensemble des séquences annotées, les UniRef sont des bases de données non redondantes et UniParc correspond aux archives.



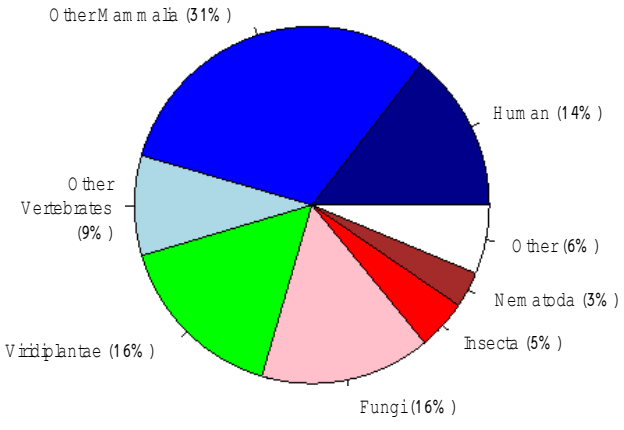


Figure 2 : Distribution taxonomique des 231434 séquences (à gauche) et des 99492 séquences eucaryotes (à droite) de la version 50.6 du 05-Sep-06 de UniProtKB/Swiss-Prot.

Les bases de données contiennent aujourd'hui plusieurs génomes entièrement séquencés. Dès le début des années 1980, il a été possible de séquencer le génome mitochondrial humain (16,5 kb, et le génome du bactériophage Lambda (40 kb, . Depuis, les techniques de séquençage ont été développées et automatisées, ce qui a permis une augmentation très importante de leur efficacité. Le premier organisme vivant dont le génome a été entièrement séquencé est la bactérie *Haemophilus influenzae* d'une taille de 1.83 Mb, rapidement suivie par *Mycoplasma genitalium* , qui possède l'un des plus petits génomes microbiens (0.58 Mb). Aujourd'hui, 411 génomes procaryotes ont été complètement séquencés et publiés, et environ 700 sont en cours d'étude. De plus, un grand nombre de génomes microbiens ont été séquencés par des projets privés, mais ces séquences ne sont pas disponibles publiquement. La levure *Saccharomyces cerevisiae* a été complètement séquencée en 1996 par un consortium international. C'est un organisme unicellulaire eucaryote modèle : elle a un petit génome (12Mb contre 3400 Mb chez l'homme), plusieurs chromosomes (16), peu de séquences répétitives d'ADN et peu d'introns. Puis la séquence complète du genome du ver *Caenorhabditis elegans* a été publiée . Ce nématode a été choisi car c'est un organisme multicellulaire simple. Aujourd'hui, 432 génomes sont connus complètement : ils correspondent à 411 organismes procaryotes, 4 animaux, 2 plantes, 9 champignons et 6 protistes. Le génome humain a été complètement séquencé en 2001 . Chacun des chromosomes humains est maintenant soumis à une analyse plus pointue incluant une identification et un référençage des séquences codantes .

I.1.2.Les banques de structures 3D de protéines

La structure tridimensionnelle de certaines séquences protéiques a pu être déterminée expérimentalement, soit avec une résolution atomique par Résonance Magnétique Nucléaire (en solution) ou par Cristallographie aux Rayons X (sur des cristaux de protéines), soit avec une résolution environ dix fois plus faible par Microscopie Electronique. Les structures tridimensionnelles connues sont répertoriées dans la Protein Data Bank (**PDB**) , dont voici les statistiques récentes (Tableau 2 ; source : http://www.rcsb.org/pdb/holdings.do).

Molecule Type						
		Proteins	Nucleic Acids	Protein/NA Complexes	Other	**Total**
Exp. Method	X-ray	30503	916	1406	28	**32854**
	NMR	4809	725	122	6	**5662**
	Electron Microscopy	91	10	29	0	**130**
	Other	75	4	3	0	**83**
	Total	**35478**	**1655**	**1560**	**34**	**38729**

Tableau 2 : Dénombrement des structures déposées dans la base de données PDB en octobre 2006.

On remarque que, par rapport au nombre de séquences disponibles, le nombre de structures tridimensionnelles déterminées est encore très faible, même si la comparaison structure/séquence n'est pas immédiate. En effet, une structure correspond parfois uniquement à un fragment d'une séquence et inversement, certaines structures contiennent plusieurs protéines.

Afin d'identifier et de classer les familles de repliements déjà connus, plusieurs équipes ont proposé des approches, soit basées sur des calculs de « distance » entre repliements, soit essentiellement construites sur un examen visuel des structures tridimensionnelles disponibles. Dans la première catégorie, la base de données **FSSP** de L. Holm et C. Sanders a été créée à partir d'un algorithme de mesure quantitative de la compacité des structures. Ce critère permet de diviser une structure en domaines de plus en plus petits. La récurrence des domaines obtenus est ensuite analysée afin de connaître la taille des domaines pouvant être retrouvés en grand nombre dans des protéines différentes. Le serveur **DALI** qui interroge FSSP permet non seulement de consulter la classification de la version la plus récente de la PDB90 (contenant les structures de moins de 90% d'identité en séquence) mais aussi de rechercher les structures les plus proches d'une structure que l'on soumet. Dans la deuxième catégorie, la base de données **SCOP** (pour Structural Classification of Proteins ; classifie en grande partie manuellement, selon une organisation hiérarchique, les 26000 structures de la PDB (Octobre 2004) en 945 repliements, 1539 super-familles et 2845 familles. Les familles rassemblent des structures clairement évolutivement proches : une famille est caractérisée par une structure et une fonction, ce qui correspond en général à plus de 30% d'identité de séquence au sein de la famille. Les super-familles rassemblent des protéines de séquences éloignées mais dont les structures et les fonctions suggèrent une origine évolutive commune.

Enfin, les repliements correspondent à des protéines présentant les mêmes éléments majeurs de structure secondaire avec un même arrangement dans l'espace et les mêmes connections topologiques. Une autre base de données construite en partie manuellement est **CATH** (pour « **C**lass, **A**rchitecture, **T**opology and **H**omologous superfamily » ; . La classification hiérarchique des structures proposée dans cette base de données est illustrée sur la Figure 3.

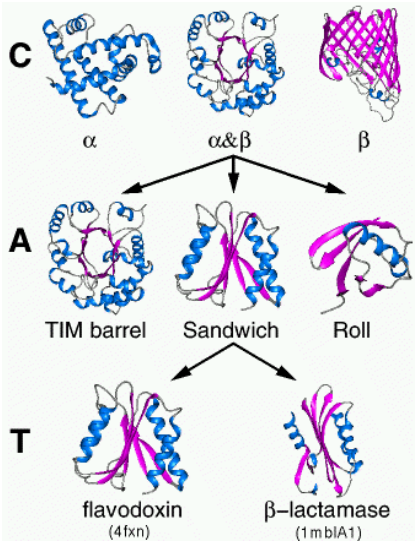


Figure 3 : Classification des repliements dans la base de données CATH (http://www.cathdb.info/cgi-bin/cath/GotoCath.pl?link=cath_info.html).

L'Architecture (40 catégories) correspond à un même arrangement des éléments de structure secondaire majeurs sans tenir compte de leurs connections topologiques, alors que la Topologie (1110 catégories) en tient tout naturellement compte. Les super-familles homologues (2147 catégories) rassemblent des protéines dont l'origine évolutive commune est prédite. Ce sont ces super-familles qui nous intéresseront tout particulièrement dans notre étude.

Certaines protéines sont constituées de l'assemblage de plusieurs régions globulaires adoptant un repliement autonome. Historiquement ces zones globulaires ont été décrites chez la papaïne et le lysozyme . Des expériences de protéolyse ménagée réalisées sur les immunoglobulines ont aussi montré que ces protéines pouvaient être clivées en plusieurs fragments stables et structurés. Pour tenir compte de cette propriété structurale, il a été défini la notion de « domaine », qui correspond à une unité de repliement autonome. Généralement, les protéines de plus de 300 acides aminés s'organisent en plusieurs domaines. Certains domaines présentent une fonction catalytique autonome, d'autres fonctions apparaissent à

l'interface entre deux domaines (cas du lysozyme). Le rapprochement entre deux domaines est rendu possible par la présence de régions non-structurées.

Plusieurs groupes ont travaillé sur le découpage en domaines structuraux des protéines de la PDB, afin d'améliorer la détection de la similarité entre domaines se trouvant au sein de protéines différentes. En particulier, la base de données 3Dee contient la définition en domaines structuraux de toutes les protéines de la PDB, revues par l'EBI et le RCSB (consortium américain « Research Collaboratory for Structural Bioinformatics »), qui ont plus de 20 résidus et qui ne sont pas des modèles théoriques. A notre connaissance, cette base de données n'a pas été mise à jour depuis 1998.

Enfin, PDB at a Glance (http://cmm.cit.nih.gov/modeling/pdb_at_a_glance.html) propose une classification des structures tridimensionnelles de la dernière version de la PDB en fonction du contexte biochimique (accès par mots-clefs).

I.2. Les méthodes de comparaison entre séquences

Les méthodes actuelles de comparaison de séquences permettent de détecter des analogies entre protéines relativement proches : ces protéines partagent en général plus de 30% de résidus identiques. Cependant, ces mêmes méthodes de comparaison de séquences restent peu sensibles lorsqu'il s'agit de détecter des ressemblances ou des homologies lointaines. Ainsi, de nombreuses séquences restent « orphelines » après l'utilisation des logiciels de recherche de similitude les plus performants. Après avoir présenté les progrès récents qui ont été faits dans ce domaine, nous montrerons comment analyser les performances des méthodes actuelles de recherche d'homologies lointaines. Enfin, nous discuterons de nouvelles approches qui pourraient permettre d'augmenter la qualité de détection de telles relations d'homologies.

I.2.1.Que cherche-t-on à détecter : la notion d'homologie

Il existe plusieurs termes permettant de nommer la ressemblance entre deux séquences biologiques. La similitude est une quantité qui mesure le pourcentage d'identité entre deux séquences. L'homologie quand à elle est une propriété qui a une connotation évolutive. Deux séquences sont dites homologues si elles possèdent un ancêtre commun. Les homologies de séquences peuvent être de deux types : orthologie ou paralogie. On parle d'orthologie si les deux séquences sont apparues après un évènement de spéciation, c'est-à-dire si le gène

ancestral est présent dans une espèce qui a divergé en deux sous-espèces conservant une copie du gène. En revanche, si les deux gènes sont séparés par un processus de duplication, ils sont dits paralogues. Pour deux séquences orthologues, les fonctions sont identiques ou proches. Pour deux séquences paralogues, cette conservation de fonction n'est pas toujours observée. En effet, après duplication, le gène n'est plus soumis à la même pression de sélection et peut alors muter et présenter de nouvelles fonctions.

L'homologie peut être déduite de la similitude. On considère qu'une similitude significative (avec un pourcentage d'identité supérieur à 40 %) est signe d'homologie sauf si les séquences présentent une faible complexité. L'inverse n'est par contre pas vrai. Une absence totale de similarité ne signifie pas absence d'homologie. L'augmentation importante du nombre de structures 3D connues a fait clairement apparaître que dans de nombreux cas, deux séquences présentant des identités de séquence de l'ordre de 20% à 40% adoptent des repliements proches et peuvent posséder des fonctions voisines. On parle alors d'homologie définie sur une base non plus génétique mais structurale. Cette homologie structurale est parfois liée à une homologie fonctionnelle.

Comme l'illustre l'étude de Mark Gerstein à partir d'une analyse sur plusieurs milliers d'alignements structuraux (Figure 4), une recherche d'homologie réalisée sur une base de séquences protéiques peut conduire à différentes interprétations en fonction de l'identité de séquence.

- Dans le cas le plus simple, une similitude forte, soit une identité supérieure à 40% et distribuée sur l'ensemble de la séquence, est trouvée. Les deux séquences appartiennent à des organismes très proches au plan phylogénétique. Vraisemblablement, les deux protéines sont homologues et présentent des structures et des fonctions identiques.
- la similitude de séquence est encore forte (identité supérieure à 40%) mais seulement sur une partie de la séquence. De plus, la conservation de certains groupes d'acides aminés semble montrer qu'une partie des fonctions associées serait présente chez les deux protéines. Ces deux séquences présentent là encore vraisemblablement des domaines structuraux homologues et des fonctions communes.
- la similitude de séquence est relativement faible (identité comprise entre 20 et 40%) sur l'ensemble de la séquence. Il est alors important de rechercher d'éventuelles analogies structurales et de déterminer si celle-ci peut correspondre à des fonctions biologiques communes. Dans cette fenêtre de similitude, on estime que les analogies structurales sont fréquentes mais n'implique pas nécessairement une similitude fonctionnelle.

- Enfin généralement très peu de résidus sont conservés (identité inférieure à 20%). Là encore, les méthodes de comparaison séquence à séquence ne permettent pas de séparer les séquences qui sont réellement homologues de celles pour lesquelles les similitudes trouvées sont fortuites (les analogies de séquences trouvées sont aléatoires). Des outils d'analyse poussés permettent parfois d'établir l'existence de similitude structurale. Dans ces cas de très faible identité, les relations fonctionnelles, si elles existent, sont généralement faibles. Ceci est illustré sur la figure 4, où la courbe noire correspond à des relations fonctionnelles fortes entre deux séquences, et la courbe bleu foncé à des relations fonctionnelles faibles. On voit que pour des pourcentages d'identité compris entre 15 et 20%, environ 20% des protéines ont des fonctions proches, alors que 40% des protéines ont des fonctions faiblement reliées et 40% des protéines n'ont pas la même fonction. Précisons que l'on parle de relation fonctionnelle faible entre deux protéines lorsque, par exemple, elles possédent le même type de repliement et de substrats, tels que le NADP et le NADPH, mais l'une opére des réactions d'oxydation et l'autre des réactions de réduction.

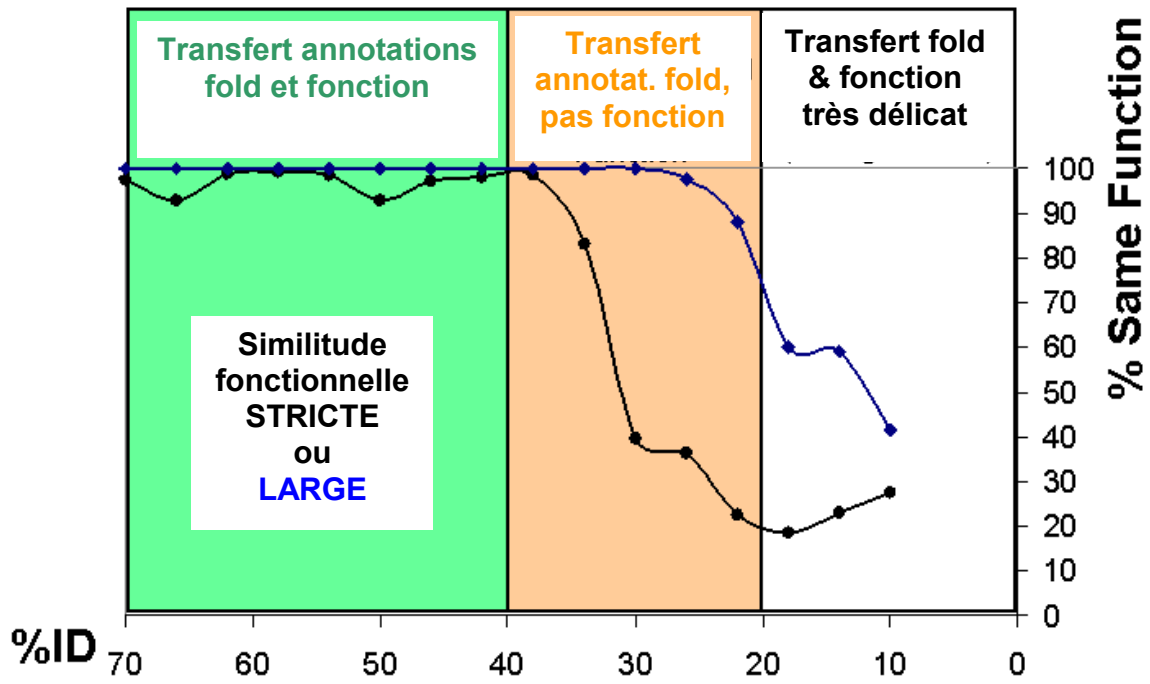


Figure 4 : Relation entre les similitudes de séquences et les similitudes fonctionnelles définies par des critères stricts (courbe noire, « narrow similarity ») ou des critères plus larges (courbe bleue, « broad similarity »). Etude réalisée par M. Gerstein sur 1000 alignements structuraux de domaines protéiques (extrait avec la permission de M. Gerstein de http://lectures.gersteinlab.org/).

Un objectif important de l'amélioration des méthodes de recherche d'homologues est de retrouver les homologues correspondant aux deux derniers cas, c'est-à-dire de faire la différence dans le bruit de fond des méthodes prédictives entre les faux négatifs - les séquences qui ne présentent pas d'analogie significative (en dessous du seuil de détection) et correspondent en fait à des protéines homologues - et les vrais négatifs - les séquences qui ne présentent pas d'analogies et ne correspondent effectivement pas à des protéines homologues.

I.2.2.Comparaison séquence à séquence

Pour aligner deux séquences entre elles, on peut chercher à optimiser le pourcentage de positions identiques dans l'alignement. Néanmoins, cette mesure ne tient pas compte de la fréquence relative des différents acides aminés qui composent les protéines. Or cette fréquence est très variable. Dans un alignement de séquence, la conservation d'un acide aminé « rare » (Tryptophane, Cystéine) n'a pas la même valeur statistique que la conservation d'un acide aminé « abondant » (Alanine, Valine). De plus, certains acides aminés présentent des structures et des fonctions chimiques voisines. Le remplacement d'un acide aminé par un autre acide aminé de fonction voisine peut être considéré comme une forme de conservation. Afin de tenir compte des caractéristiques propres à chaque acide aminé, des matrices de scores ont été développées.

I.1.1.1 Les matrices de scores

Ces matrices permettent de réaliser une pondération des remplacements d'un acide aminé par un autre selon divers critères. Il s'agit donc d'attribuer des scores associés aux 400 changements possibles (20 X 20) entre acides aminés. Plusieurs matrices ont été proposées. Certaines sont basées sur les caractéristiques physico-chimiques des acides aminés, d'autres ont été construites à partir de l'analyse des statistiques des substitutions au sein d'un ensemble de familles de séquences alignées. Ces dernières, telles que les matrices PAM ou BLOSUM, sont couramment utilisées et nous mentionnons ici succinctement leurs caractéristiques.

(i) Les matrices PAM

Margaret Dayhoff et ses collègues ont établi une méthode permettant d'estimer la probabilité qu'un acide aminé i soit remplacé par un acide aminé j à une position donnée . Cette estimation repose sur l'analyse d'alignements de séquences proches au sein desquelles il est peu probable qu'une mutation de A en B résulte de mutations successives A>X>Y>B.

Comme un nombre relativement faible d'alignements a été analysé, la matrice constituée des fréquences de mutations observées comporte des 0 et des 1. Elle a été complétée en calculant, dans le cas où une mutation particulière n'a pas été observée, le rapport entre le nombre de mutations auquel ce type d'acide aminé est soumis et le nombre total d'acides aminés de ce type présent dans les alignements. Les matrices de Dayhoff utilisées aujourd'hui ont été obtenues à partir de l'analyse de 71 alignements globaux de protéines de fonctions identiques (environ 1300 séquences). Elles dépendent des différences de séquences, fréquences et mutabilités observées au sein de ces 1300 séquences. Chaque matrice donne les probabilités de mutations après un temps d'évolution donné. L'analyse de ces matrices permet de construire un arbre phylogénétique : une recherche des séquences ancestrales pour chacun des noeuds internes de l'arbre est effectuée en utilisant le principe du maximum de parcimonie qui consiste à réaliser un nombre minimum de changements pour passer d'une séquence à l'autre.

M. Dayhoff et ses collègues utilisent un paramètre δ qui correspond à la proportion d'acides aminés ayant muté après l'intervalle d'évolution représenté par la matrice de probabilité PAM (Percentage of Acceptable point Mutations per 10^8 years). Lorsque les paramètres de l'analyse sont ajustés afin que δ soit égal à 1 (1 mutation observée pour 100 sites), on obtient une matrice nommée PAM1. Cette matrice peut être utilisée pour générer des matrices correspondant à un temps d'évolution plus lent en la multipliant par elle-même de manière répétée. Les matrices PAMn sont alors définies comme $(PAM1)^n$. Lorsque l'on compare des séquences dont on sait qu'elles sont reliées, la matrice PAM200 est appropriée. Lorsque l'on recherche des homologues dans une base de données, PAM120 paraît être le meilleur compromis. Enfin, lorsque l'on utilise une méthode d'alignement local, les matrices PAM40, PAM120 et PAM250 sont intéressantes : elles permettent tour à tour de détecter des séquences courtes et très homologues puis longues et moins homologues.

Plus récemment, le groupe de J.M. Thornton a obtenu une matrice de substitution construite de manière analogue aux matrices PAM, mais basée sur l'analyse de 2621 familles de séquences de la base de données SWISSPROT version 15.0 . Les principales différences entre cette nouvelle matrice nommée PET91 et les matrices de Dayhoff concernent les substitutions qui étaient mal représentées dans l'étude de 1978. PET91 apparaît ainsi comme une mise à jour des matrices de Dayhoff.

(ii) Les matrices BLOSUM

Henikoff et Henikoff ont tenté de dériver des probabilités de mutations à partir de séquences protéiques plus éloignées. Tout d'abord, une base de données d'alignements multiples de courtes régions sans gap a été obtenue. L'hypothèse sous-jacente est que ces blocs correspondent à des éléments bien conservés au sein des structures tridimensionnelles des protéines correspondantes. Puis, au sein de chaque alignement, les séquences ont été regroupées en familles partageant plus qu'un certain pourcentage seuil d'identité. Les fréquences de substitution pour toutes les paires d'acides aminés ont alors été calculées et utilisées pour obtenir une matrice BLOSUM (BLOcks SUbstitution Matrix). Différentes matrices ont été calculées en faisant varier le pourcentage d'identité seuil. Par exemple, la matrice BLOSUM 62 a été obtenue en utilisant un seuil de 62% d'identité (Figure 5). Si deux séquences d'un même alignement initial ont plus de 62% d'homologie, alors elles sont représentées une seule fois dans l'alignement servant à construire la matrice BLOSUM62.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Figure 5 : Matrice de substitution BLOSUM62. Les scores positifs en rose indiquent les acides aminés considérés comme similaires. Les scores de la diagonale correspondent aux valeurs attribuées en cas d'identité.

(iii) Les matrices dérivées des alignements structuraux

La comparaison des structures 3D des séquences à aligner permet d'améliorer l'exactitude de l'alignement, en particulier lorsque l'homologie entre les séquences est faible. L'analyse d'alignements structuraux devrait donc donner les meilleures matrices de substitutions. Ainsi, Risler et al. a dérivé des fréquences de substitution à partir de 32

séquences protéiques réparties au sein de 11 alignements. De la même façon, Overington et al. a aligné 7 familles pour lesquelles plus de 3 protéines avaient une structure 3D connue et en a dérivé une série de matrices de substitution. Enfin, Bowie et al. (1991) a dérivé des tables de substitution pour différents environnements en acides aminés et différents éléments de structure secondaire.

(iv) Comparaison des différentes matrices

Le consensus général est que les matrices dérivées de l'observation des données de substitution (i.e. les matrices de Dayhoff ou BLOSUM) sont meilleures que les matrices fondées sur le code génétique ou les propriétés physiques. Les matrices BLOSUM aux PAM ont été comparées à des matrices issues d'alignements structuraux . Il a été conclu que la matrice BLOSUM62 est en général la plus efficace. On peut néanmoins supposer que orsque davantage de structures 3D seront disponibles, les tables de substitutions dérivées des comparaisons structurales pourront fournir des matrices plus fiables.

I.1.1.2 La prise en compte des insertions dans les alignements par paires

La comparaison locale de deux séquences (nucléiques ou protéiques) repose sur l'hypothèse de microévolution par mutations ponctuelles (remplacement d'un acide aminé ou nucléotide par un autre, suppression ou insertion). Pour évaluer la qualité d'un alignement, un score élémentaire issu de la matrice de substitution est calculé à chaque position alignée et une pénalité est également calculée en cas d'insertions. La valeur attribuée aux pénalités d'insertions est généralement calculée avec une fonction affine associant une pénalité d'ouverture de l'insertion et une pénalité d'extension augmentant linéairement avec la longueur de l'insertion. L'avantage de ce calcul est sa simplicité mais la modélisation de la pénalité à attribuer aux insertions fait encore l'objet d'optimisation, en particulier dans le cas de l'alignement de séquences très divergentes.

I.1.1.3 Les algorithmes d'alignement

Le nombre de façons d'aligner deux séquences entre elles en tenant compte des insertions est très important. Pour deux séquences de longueur M et N, le nombre d'alignements possibles se calcule avec l'expression :

$$Nb = 2^M 2^N$$

Par exemple, pour deux séquences de 20 résidus, le nombre d'alignements possibles est de $2^{(40)}$ soit 137 milliards de combinaisons.

(i) La programmation dynamique :

Les méthodes de programmation dynamique sont des approches algorithmiques développées par le mathématicien Richard Bellman visant à réduire la complexité d'un problème combinatoire complexe. Appliqué au problème des alignements de séquence, les méthodes de programmation dynamique construisent un alignement optimal de sous-séquences de plus en plus longues en utilisant les scores obtenus pour les sous-séquences. Les méthodes de programmation dynamique les plus couramment utilisées pour l'alignement de séquences sont les algorithmes de Needleman et Wunsch (1970) et Smith et Waterman (1981).

(ii) Needleman et Wunsch

L'algorithme de Needleman et Wunsch a été développé pour réaliser l'alignement global de deux séquences protéiques . Cet algorithme s'organise en trois étapes. Tout d'abord, une matrice de similarité dont les deux dimensions correspondent aux deux séquences à aligner est remplie en attribuant à chaque élément la valeur extraite de la matrice de substitution (BLOSUM62 par exemple).

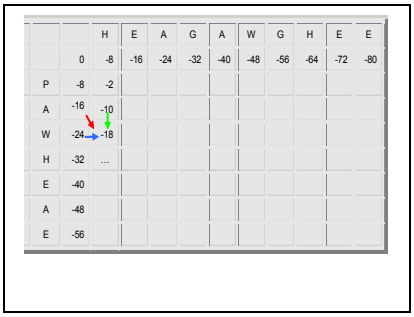


*Figure 6 : Construction de la matrice des scores d'alignement entre les deux séquences «HEAGAWGHEE » et « PAWHEAE » calculée à partir d'une matrice de similarité basée sur les scores de substitution BLOSUM62 et un coût d'insertion (Winsert) constant de -8. Pour chaque cellule, le score maximal obtenu à partir des trois flèches de couleur est attribué à cette cellule. Dans l'exemple la flèche rouge correspond à une absence d'insertion et le score associé se calcule par Si,j = Si-1,j-1+Sblosum62(W,H) = -16 -2 = -18, la flèche verte correspond à une insertion dans la séquence 1 et le score associé est Si,j = Si,j-k+k*Winsert = -10 - 8 = -18 (k=1), et la flèche bleue correspond à une insertion dans la séquence 2 et le score associé est Si,j = Si-l,j+l*Winsert = -24-8 = -32 (l=1). Dans la Figure 7 ci-dessous, seuls les parcours passant par les flèches rouges et vertes sont conservés car ils sont associés aux trajectoires ayant fourni les scores maxima.*

Cette matrice est ensuite utilisée pour construire une seconde matrice (regroupant les scores d'alignements) remplie, à chaque position, par le score maximal d'un alignement qui se terminerait à cet élément (règle de récurrence simple présentée en Figure 6. Enfin, la lecture

des scores les plus élevés dans la matrice des scores d'alignements, parcourue en sens inverse, permet d'identifier l'alignement global optimal entre les deux séquences (flèches vertes de la Figure 7).

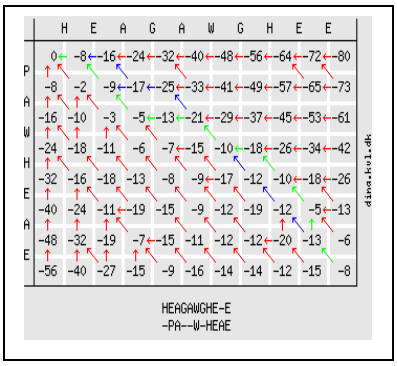


Figure 7 : Lecture de la matrice de scores d'alignement afin d'identifier l'alignement optimal par l'algorithme de Needleman et Wunsch. Les flèches indiquent à partir de quelle cellule le score d'alignement a été obtenu à l'étape de construction de la matrice (Figure 6). Les flèches vertes indiquent le parcours produisant le score optimal parmi les flèches rouges illustrant les sous-parcours optimaux. Les flèches bleues indiquent les parcours produisant les mêmes scores que le parcours optimal. L'alignement optimal correspondant au parcours vert est présenté en dessous.

(iii) L'algorithme de Smith-Waterman

L'algorithme de Smith et Waterman permet d'identifier l'alignement optimal local et non global entre deux séquences. La procédure est directement inspirée de celle de Needleman et Wunsch. La principale différence vient du fait que n'importe quel élément de la matrice de scores d'alignements peut être considéré comme point de départ pour l'identification de la trajectoire maximisant les scores d'alignements (Figure 8). Si le score d'alignement devient inférieur à zéro, la case est réinitialisée à zéro et peut être considérée comme un nouveau point de départ.

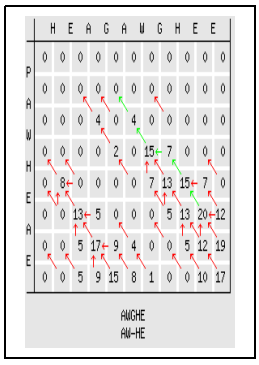


Figure 8 : Lecture de la matrice de scores d'alignement afin d'identifier l'alignement optimal par l'algorithme de Smith et Waterman. Les flèches indiquent de quelle cellule le

score d'alignement a été obtenu à l'étape de construction de la matrice. Les flèches vertes indiquent le parcours produisant le score optimal parmi les flèches rouges illustrant les sous-parcours optimaux. L'alignement optimal correspondant au parcours vert est présenté en dessous.

(iv) Développement des méthodes heuristiques

Les méthodes de programmation dynamique permettent l'obtention d'un alignement optimal entre deux séquences, mais elles présentent un coût en temps de calcul et en mémoire important. Ces méthodes sont donc adaptées pour aligner un nombre limité de séquences mais pas pour comparer une séquence avec l'ensemble des séquences présentes dans les bases de données. Le but des méthodes heuristiques est de calculer à moindre coût des alignements pas nécessairement optimaux mais de qualité suffisante pour établir des relations d'homologies entre les séquences. Les méthodes heuristiques les plus utilisées sont les algorithmes FASTA et BLAST.

(v) l'algorithme FASTA

Le logiciel FASTA, développé par Pearson et Lipman , procède en quatre étapes.

(1) Comparaison de la position et de la nature de segments de longueur L entre deux séquences afin de repérer les régions les plus denses en identités partagées.

(2) Evaluation des 10 régions présentant les plus hauts scores de similarité calculés à partir des matrices de substitution. Cette étape correspond à une recherche de similitude sans insertions. Le score **init1** est attribué à la région ayant le plus fort score parmi les 10 analysées (**initn** est le score pour les autres régions).

(3) Jonction des régions précédentes, s'il en existe au moins deux et si chacune d'elles possède un score supérieur à un seuil donné. Ce seuil correspond à un score moyen attendu pour des régions non apparentées. Les régions initiales sont réunies à chaque fois que leur score diminué d'une pénalité de jonction est supérieur ou égal au score init1. Cette étape permet d'éliminer les segments peu probables parmi ceux définis à l'étape précédente.

(4) Alignement optimal (par programmation dynamique) des deux séquences en considérant uniquement les régions définies à l'étape précédente avec calcul d'un score (**opt**).

Le programme calcule un z-score qui correspond au score maximum attendu normalisé.

(vi) L'algorithme BLAST

BLAST (Basic Local Alignment Search Tool) est un programme de recherche de similarité qui a été développé au NCBI (). Son succès repose sur une grande efficacité

algorithmique associée à une fiabilité satisfaisante. Le principe du programme est de découper la séquence requête en mots élémentaires (typiquement de 3 acides aminés) et de rechercher tous les mots de la base de données qui s'alignent avec ce mot au dessus d'un score seuil (appelés les HSP pour High-scoring segment pairs). Les HSPs sont ensuite agrégés pour former des fragments d'alignements locaux possédant des scores très élevés (nommés MSP pour Maximal-scoring Segment Pair). Ces MSPs sont alors étendus jusqu'à ce que le score d'alignement descende d'une quantité seuil. L'alignement local produit est alors évalué et un score lui est associé.

Pour évaluer la validité statistique d'un alignement, le programme BLAST s'appuie sur une expression analytique qui permet de traduire directement le score de l'alignement en probabilité que cet alignement soit significatif. La distribution des scores d'alignements obtenus par la comparaison d'une séquence avec l'ensemble des séquences d'une base de données est supposée suivre une loi de distribution dite de la valeur extrême. Le paramètre nommé « e-value » caractérisant la probabilité qu'un alignement soit significatif peut alors s'écrire :

$$Evalue = K \cdot L_1 \cdot L_2 \cdot e^{-\lambda \cdot Score}$$

Dans cette expression L_1 et L_2 sont les longueurs des deux séquences à aligner, K et λ des paramètres de normalisation ajustés en fonction de la taille de la base de données, de la taille des séquences et de la prise en compte ou non des insertions. La valeur de la e-value caractérise le nombre de séquences non-homologues susceptibles d'être détectées avec ce score. Plus la e-value est faible, plus l'alignement est donc supposé rendre compte d'une relation d'homologie entre deux séquences.

I.2.3.Les méthodes d'alignements multiples

Partant d'une séquence d'intérêt, nous avons vu qu'il est possible d'identifier un ensemble de séquences homologues par des comparaisons par paires avec les séquences des bases de données. L'alignement multiple de ces séquences homologues a ensuite pour objectif d'agencer en colonne les acides aminés qui possèdent la même histoire évolutive. L'obtention de ces alignements multiples constitue une étape essentielle de l'analyse bioinformatique car elle permet de mettre en évidence les positions importantes pour la structure et/ou la fonction. Dans le contexte de la détection d'homologie lointaine et de la prédiction de structure, l'optimisation des alignements multiples revêt donc une importance cruciale.

I.1.1.4 . Alignements multiples et profiles : l'approche de PSI-BLAST

L'approche la plus simple et la plus rapide pour construire un alignement multiple consiste à utiliser la séquence d'intérêt comme référence de l'alignement et à ajouter les séquences homologues une à une, en intégrant les insertions telles qu'elles ont été définies par l'alignement par pair avec la séquence d'intérêt. Cette stratégie génère généralement un grand nombre d'insertions qui rend les alignements multiples difficile à exploiter. Néanmoins, en première approximation cette stratégie est efficace. Le logiciel PSI-BLAST, qui a révolutionné la détection des homologies à faible identité de séquence, utilise ce type de méthode pour générer un alignement multiple après une première itération de l'algorithme BLAST. La distribution des acides aminés et des insertions dans chaque colonne de l'alignement multiple permet d'extraire une fréquence d'occurrence pour chaque position qui peut être traduite en termes de scores de probabilité. La table qui ordonne l'ensemble de ces fréquences est appelée profil, ou, dans le cas de PSI-BLAST, PSSM (pour position-specific score matrix). L'exemple de la Figure 9 illustre la structure d'un profil similaire à ceux générés par le logiciel PSI-BLAST.

POS	PROBE	CONSENSUS		PROFILE																		
				A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	+/-
1	E G V L		V	3	-2	3	4	0	4	-1	3	-1	4	4	1	1	1	-2	1	2	6	-6	-2	9
2	L L S P		L	2	-2	-2	-1	3	0	-1	3	-1	6	5	-1	3	0	-1	3	1	4	1	-1	9
3	V V V V		V	2	2	-2	-2	2	2	-3	11	-2	8	6	-2	1	-2	-2	0	2	15	-9	-1	9
4	K E A T		A	6	-2	5	6	-5	4	1	0	5	-2	0	3	3	3	1	3	6	0	-6	-4	9
5	A P L P		P	6	-1	0	1	-2	2	0	1	0	2	2	0	8	2	0	2	2	3	-5	-4	9
6	G G G G		G	7	1	7	5	-6	15	-1	-3	0	-4	-3	4	3	2	-3	6	4	2	-11	-7	9
7	S S Q E		D	4	-1	7	7	-6	7	2	-2	2	-3	-2	4	3	6	1	6	2	-1	-6	-5	9
8	S S T P		S	4	4	2	2	-4	4	-1	0	2	-3	-2	2	7	0	1	10	6	0	-2	-4	9
9	V L V A		V	5	0	-1	-1	3	1	-2	7	-2	7	6	-1	1	-1	-3	0	2	10	-5	-1	9
10	K R R S		R	0	-1	1	1	-5	0	2	-2	8	-3	1	3	3	3	10	5	1	-2	7	-5	9
11	M L I I		I	0	-2	-3	-2	7	-3	-3	11	-1	11	10	-2	-2	-1	-2	-2	1	9	-3	1	9
12	S S T S		S	4	6	2	2	-3	5	-1	0	2	-3	-2	3	4	-1	1	12	6	0	0	-4	9
13	C C C C		C	3	15	-5	-5	-1	2	-1	3	-5	-8	-6	-3	1	-6	-3	7	3	3	-13	10	9
14	K S Q R		K	1	-2	3	3	-6	1	3	-2	7	-3	0	3	3	5	7	4	1	-2	2	-5	9
15	A A G S		A	10	3	4	3	-5	8	-1	-1	1	-2	-1	3	4	1	-2	7	4	2	-6	-4	9
16	T S D S		S	4	3	5	4	-5	6	0	0	2	-3	-2	4	3	1	1	9	6	0	-3	-4	9
17	G G S Q		G	5	1	6	5	-6	9	1	-2	1	-3	-2	4	3	4	0	6	3	0	-6	-6	9
18	Y F L S		F	-1	2	-4	-3	9	-3	0	4	-3	6	3	-1	-3	-3	-3	1	-1	2	7	7	9
19	T T R L		T	1	-2	0	1	0	0	0	2	2	2	3	1	1	1	3	1	7	2	1	-2	9
20	F F . L		F	-2	-3	-6	-4	10	-4	-1	6	-4	9	6	-3	-4	-4	-3	-2	-1	3	7	8	4
21	S S . D		S	3	2	5	4	-4	5	0	-1	2	-3	-2	4	3	1	1	8	2	-1	-2	-3	4
22	S . . S		S	2	3	1	1	-2	3	-1	0	1	-2	-1	2	2	0	1	8	2	0	1	-2	4
23	. . . G		G	2	0	2	1	-2	4	0	0	0	-1	-1	1	1	1	-1	2	1	1	-3	-2	4
24	. . . D		D	1	-1	4	3	-2	2	1	0	1	-1	-1	2	1	2	0	1	1	0	-3	-1	4
25	. . . G		G	2	0	2	1	-2	4	0	0	0	-1	-1	1	1	1	-1	2	1	1	-3	-2	4
26	. A G N		A	6	0	4	3	-4	6	1	-1	1	-2	-1	5	2	2	-1	3	3	1	-5	-3	4
27	Y N Y T		Y	0	5	0	-1	5	-1	2	1	-1	0	-1	4	-3	-2	-2	0	3	0	3	6	4
28	E D D Y		D	2	-2	9	8	-3	3	4	-1	1	-3	-2	5	-1	4	-1	1	1	-1	-6	0	9
29	L M A L		L	3	-5	-3	-1	6	-1	-2	6	-1	10	10	-2	0	0	-2	-1	0	6	-1	0	9
30	Y N A W		N	4	1	3	2	0	2	3	-1	1	-1	-1	8	0	1	-1	2	1	-1	-1	2	9
.	.													.										
.	.													.										
.	.													.										
48	S G N S		S	4	3	5	3	-4	7	0	-2	2	-4	-3	6	3	1	0	10	3	0	-2	-4	9
49	S S N Y		S	2	5	2	1	1	2	1	0	1	-2	-2	5	1	-1	0	8	1	-1	3	1	9

Figure 9 : Représentation d'un profil paramétrant un alignement de 4 séquences présenté en vertical à gauche. Pour chaque position un score de substitution différent est calculé pour chaque acide aminé et pour les insertions .

Le profil ainsi défini peut être utilisé pour une seconde itération dans laquelle les matrices BLOSUM ou PAM sont substituées par le PSSM. Plusieurs itérations peuvent ainsi permettre d'affiner progressivement le profil et d'augmenter considérablement la sensibilité de la méthode de détection. Ces itérations peuvent être effectuées jusqu'à la convergence, c'est-à-dire jusqu'à ce qu'aucune nouvelle séquence ne soit détectée. L'un des dangers de l'approche itérative reste la possible intégration d'un faux positif au groupe de séquences homologues lors d'une des étapes d'itération. Ce faux positif est alors susceptible de biaiser l'ensemble du profil aux itérations ultérieures. Plusieurs optimisations ont été proposées pour réduire le danger potentiel de l'intégration de ces faux-positifs . De plus, l'approche itérative telle qu'elle vient d'être décrite présente le risque, lorsque le nombre de séquences est faible, de mal estimer les probabilités d'occurrence de certains acides aminés et ainsi de fausser le profil associé à la famille de séquences homologues. Par exemple, l'observation d'une position exclusivement occupée par une isoleucine devrait laisser la possibilité que d'autres hydrophobes tels que la leucine ou la valine soient également probables. Pour améliorer un profil, il est possible d'enrichir les fréquences observées dans l'alignement par la connaissance que l'on a, a priori, des relations entre acides aminés. Cette connaissance a priori peut être intégrée de façon très détaillée mais complexe en utilisant les distributions de probabilités proposées dans les mélanges de Dirichlet qui dépendent des différents contextes observés dans une colonne d'un alignement . Plus simplement, les fréquences d'occurrence observées peuvent être corrigées par la méthode des « pseudo-count » comme c'est le cas dans PSI-BLAST. Dans cette approche on ajoute une contribution variable des scores des matrices BLOSUM et PAM aux fréquences observées. L'importance de ces scores de « connaissance a priori » est pondérée en fonction de la richesse d'information déjà contenue dans l'alignement multiple.

I.1.1.5. Les approches HMM

Une généralisation de la notion de profil a été développée en utilisant le formalisme des Chaînes de Markov Cachées (HMM) . Le formalisme associé aux HMM fournit un ensemble d'outils statistiques très performants pour manipuler et évaluer la vraisemblance d'un alignement. Les méthodes HMM sont très utilisées dans le traitement des séquences à grande échelle, dans la constitution et l'interrogation des bases de données de domaines telles que PFAM et SMART , et également pour la détection et l'alignement des homologues lointains.

Le formalisme HMM permet de générer un modèle statistique d'un alignement multiple dans lequel l'apparition des acides aminés dans l'alignement suit un processus stochastique de Markov (la probabilité de l'état n dépend uniquement de l'état n-1). Un alignement multiple peut ainsi être modélisé par une chaîne d'éléments qui possèdent 3 états (M pour une position alignée, I pour une insertion, D pour une délétion) avec des probabilités d'émission et de transition attribuées à et entre chacun de ces états (Figure 10).

Les modèles HMM fournissent une flexibilité accrue par rapport aux profils en autorisant les états de délétions en plus des états d'insertions. Les probabilités qui sous-tendent un alignement sont inconnues (variables « cachées ») et le premier objectif est de les estimer à partir des fréquences d'occurrence observées à chaque position. Comme pour les profils présentés précédemment cette information est enrichie par la connaissance a priori des probabilités d'occurrence des acides aminés en fonction des contextes et sont typiquement intégrés en utilisant les mélanges de Dirichlet.

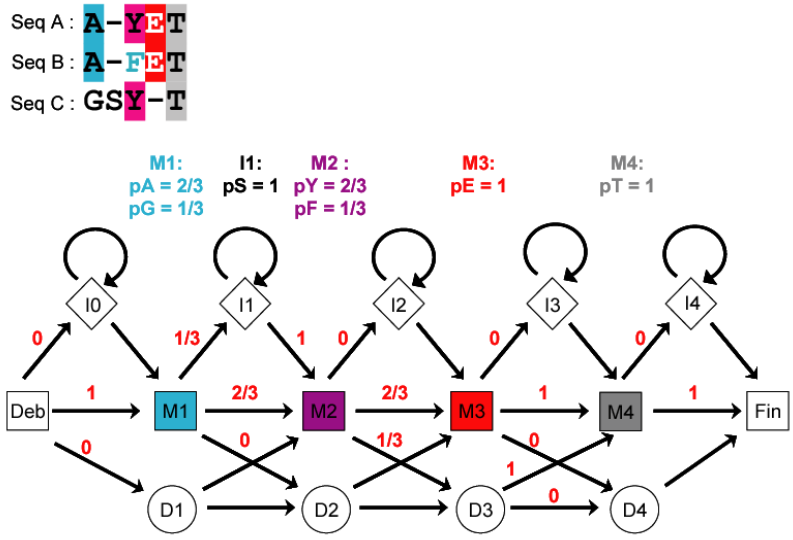


Figure 10 : Architecture simplifiée d'un modèle HMM (HMMER plan 7) et exemple de paramétrisation. L'alignement des trois séquences A,B, et C en haut peut être représenté par une chaîne à 4 états. Les probabilités de transitions entre ces états déduites de l'alignement multiple sont indiquées en rouge au dessus de chaque flèche. Les probabilités d'émission de chaque acide aminé au sein de chaque état sont indiquées en couleur au dessus du modèle.

Le modèle HMM ainsi paramétré peut être utilisé pour reconnaître les séquences susceptibles d'être reliées aux séquences de l'alignement (Figure 11). L'algorithme de Viterbi, de façon similaire aux algorithmes de programmation dynamique, permet d'identifier la trajectoire la plus probable au sein du modèle HMM.

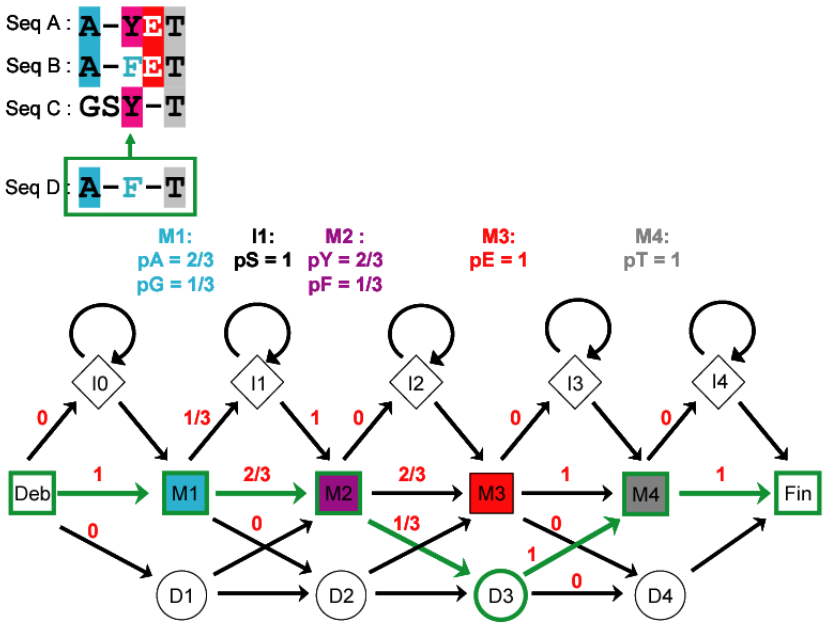


Figure 11 : Identification de la trajectoire pour que le HMM génère la séquence D alignée de façon optimale par l'algorithme de Viterbi. La trajectoire permettant de maximiser des probabilités lors du parcours du HMM est indiquée en vert. L'alignement optimal correspondant pour la séquence D est indiqué en haut.

Formellement, une séquence n'est pas alignée sur un HMM. Ce que l'on mesure, c'est la probabilité qu'un HMM donné puisse générer la séquence alignée de façon optimale. *A priori*, il serait possible d'utiliser des programmes tels que HMMER pour construire les alignements multiples de façon itérative comme le fait PSI-BLAST. Néanmoins, cette démarche est peu répandue, pour des raisons de temps de calcul important et du fait de la facilité d'utilisation de PSI-BLAST en serveur. On peut noter tout de même que le serveur MPI ToolKit du Max Planck à Tűbingen propose une utilisation couplée de HMMER et de PSI-BLAST pour la construction itérative de profils à partir de bases de séquences pré-filtrées afin d'accélérer le traitement des requêtes .

La traduction des alignements multiples en profils modélisés sous forme de matrices ou de HMM a considérablement augmenté la sensibilité des méthodes de détection de séquences homologues et a révolutionné les pratiques d'analyses bioinformatiques au cours des années 90. Néanmoins, dans l'ensemble des applications que nous venons de voir, les algorithmes employés pour générer l'alignement multiple lui-même peuvent être à l'origine d'erreurs d'alignement dont il est important de comprendre l'origine.

I.1.1.6. Evaluation de la qualité d'un alignement multiple

(i) Fonction objective d'un alignement multiple

Le meilleur alignement multiple pour un ensemble de séquences correspond à celui qui amène dans chaque colonne le maximum d'acides aminés similaires. La qualité d'un alignement peut être mesurée par une fonction mathématique appelée fonction objective. Par exemple, pour trois séquences alignées A, B, C, la fonction objective « scores des paires » calcule la somme des scores de substitution entre les couples (A,B), (B,C) et (A,C) à chaque position. Bien que cette métrique pose des problèmes dans son interprétation évolutive, elle demeure largement la plus utilisée car relativement fiable et rapide à calculer

(ii) ***La question combinatoire***

Le choix d'une fonction objective fiable ne résout qu'une partie du problème. L'analyse combinatoire nécessaire pour identifier le meilleur alignement parmi l'ensemble des alignements possibles devient rapidement considérable avec le nombre de séquences à aligner. Pour aligner deux séquences par programmation dynamique d'une longueur L de 100 acides aminés, il faut effectuer 30.000 opérations de comparaison. La même démarche de programmation dynamique naïvement généralisée à six séquences requiert déjà 2^{48} opérations car la complexité augmente en $O(2^{nbseq}L^{nbseq})$. Il s'agit d'un problème qui n'est pas soluble de façon exacte par les méthodes algorithmiques. Les progrès effectués par les méthodes d'alignements se sont donc concentrés sur des stratégies heuristiques permettant de réduire la complexité de la recherche tout en atteignant une précision maximale. Le Tableau 3 rassemble une sélection d'algorithmes d'alignement multiple avec quelques unes de leurs caractéristiques importantes. Pour établir les performances de ces algorithmes, il a été nécessaire de disposer d'alignements de références « corrects ». Construits et optimisés sur la base des superpositions entre structures tridimensionnelles, les bases de données d'alignements de référence jouent un rôle prépondérant dans la validation des méthodes d'alignement. Lors de la présentation des algorithmes, il est donc important de préciser les caractéristiques de bases de données qui sont utilisées pour les évaluer.

(iii) Bases de données d'alignements multiples de référence

Une des composantes essentielles du processus d'amélioration des méthodes d'alignements multiples a été le développement conjoint de bases de données permettant d'évaluer de façon objective les progrès réalisés. Cette démarche a été initiée par la construction de la base de données d'alignements multiples BAliBASE (~ 217 alignements) qui s'organise en différentes catégories d'alignements multiples . Chaque catégorie permet d'évaluer les performances des méthodes d'alignement pour un grand nombre de cas de

figure, tels que des séquences plus ou moins divergentes, l'existence d'insertions ou d'extensions de grandes tailles, des séquences possédant des motifs répétés, etc …. Plus récemment d'autres bases de données ont également été publiées telles que SABmark ou PREFAB . SABmark contient un grand nombre d'alignements multiples (~ 634 alignements) dont les alignements de référence ont été obtenus à partir de la superposition structurale des domaines de la base de données SCOP. Elle s'intéresse exclusivement à l'alignement des séquences divergentes et est organisée en deux catégories, « twighlight » (comprenant des séquences très divergentes dont les identités sont inférieures à 20%) et superfamille (directement relié à la définition de superfamille et de famille de repliement employé dans SCOP). PREFAB (~ 1932 alignements) a été publié simultanément au développement du programme MUSCLE (cf plus loin) et s'appuie comme SABmark sur des alignements structuraux générés de façon automatique. Enfin, la base de données d'alignements de structures homologues HOMSTRAD , traduite en alignements de séquences, est également fréquemment utilisée pour évaluer les performances des méthodes d'alignements multiples en fonction du pourcentage d'identité moyen entre les séquences de l'alignement.

	Méthode	VITESSE	PRECISION	REFERENCE
CLUSTALW	Matrice	++++	+	
DIALIGN	Matrice	+	-	
T-COFFEE	Matrice	-	+++	
MUSCLE	Matrice	++	+++	
PROBCONS	HMM	+	++++	
MAFFT	Matrice	+++	++	
MUMMALS	HMM	-	++++	
KALIGN	Matrice	++++	+++	
SPEM	Matrice	--	++++	

Tableau 3 : Table récapitulative des programmes d'alignements multiples mentionnés dans ce chapitre.

I.1.1.7. Stratégies d'alignement multiples et règles heuristiques

Pour présenter un état des lieux du développement des méthodes d'alignement multiple, il est important de préciser à quel niveau sont appliquées les règles heuristiques. Le schéma (Figure 12) indique de façon générique les différentes étapes de la construction d'un alignement multiple et chaque cadre jaune correspond aux différentes méthodes que nous allons discuter.

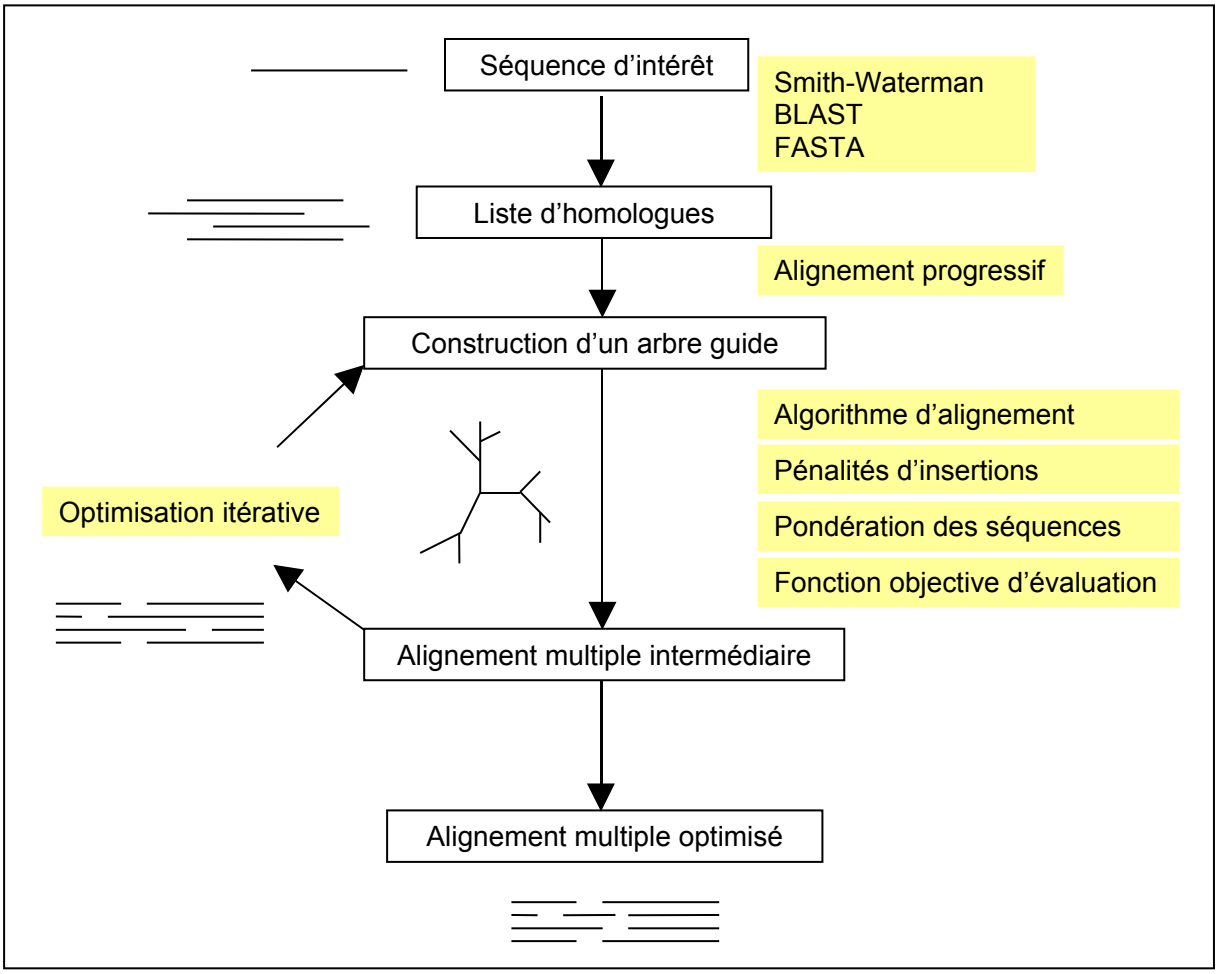


Figure 12 : Schéma simplifié des différentes étapes de calcul d'un alignement multiple. Les cadres blancs indiquent les objectifs et les cadres jaunes les différentes méthodes appliquées pour parvenir à ces objectifs.

(i) Alignements progressifs.

Une stratégie heuristique utilisée par la plupart des méthodes consiste à construire l'alignement de façon progressive. Pour aligner N séquences, on cherche à réaliser N-1 alignements par paires et à agréger ces alignements de façon hiérarchique en se guidant avec l'arbre phylogénétique déduit des alignements par paires. Les séquences les plus similaires sont alignées entre elles en priorité. En suivant cette procédure on rationalise la distribution

des insertions au sein de l'alignement multiple. Dès qu'une insertion est ajoutée dans un alignement, elle se propage dans la suite du processus d'alignement et ne sera pas éliminée.

Pour fusionner les différents alignements, les scores contenus dans leurs profils peuvent être utilisés dans des algorithmes de programmation dynamique similaires à ceux décrits précédemment pour le cas d'alignement entre paire de séquences. De façon progressive, les alignements peuvent ainsi être agrégés par alignement de séquences à un profil ou par alignement des profils entre eux. Nous reviendrons plus tard sur les questions posées par l'alignement de deux profils.

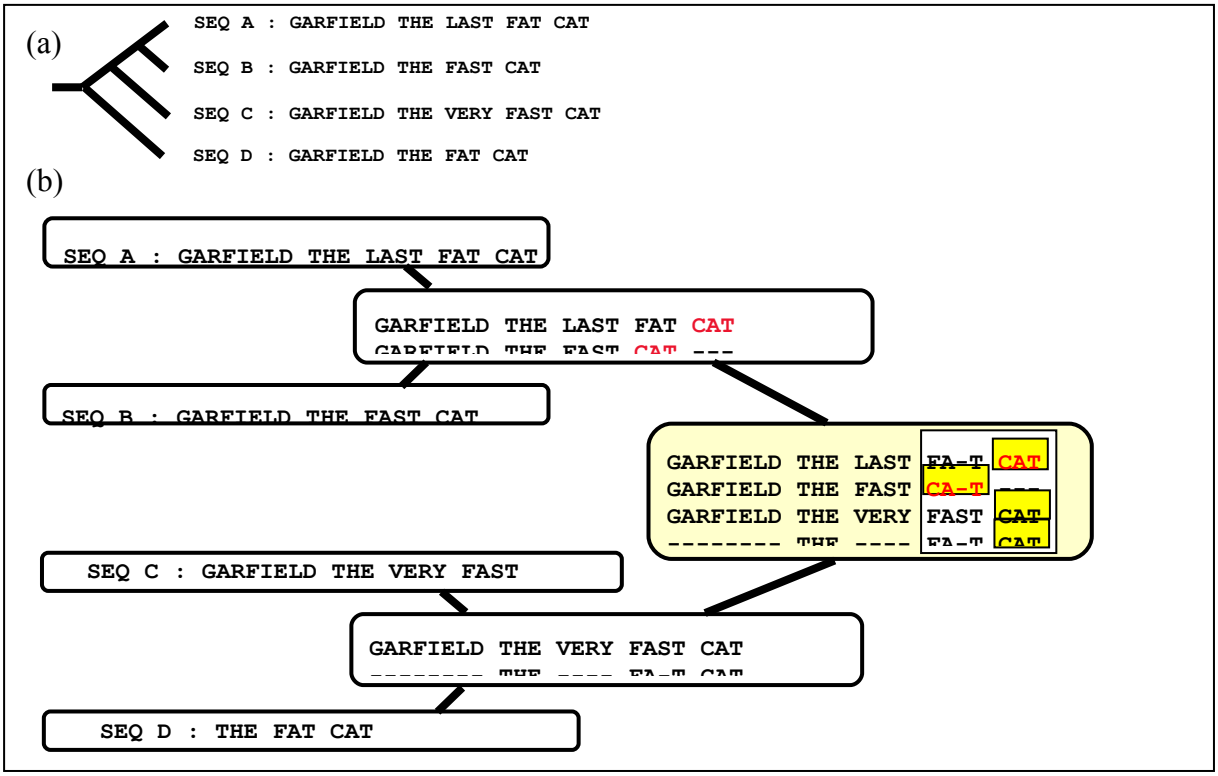


Figure 13 : (a) Exemple de 4 séquences avec l'arbre phylogénétique guide correspondant. L'arbre définit l'ordre d'alignement suivant 1. A et B, 2. C et D, 3. (A,B) et (C,D). (b) L'alignement final comporte une erreur indiquée en rouge, introduite lors de l'alignement entre A et B et qui ne peut plus être corrigée dans la suite. .

(ii) Stratégies pour pallier aux limitations des alignements progressifs

La stratégie d'alignement progressif utilisée dans un algorithme tel que Clustalw comporte des limitations et peut conduire à des erreurs systématiques dès que les séquences sont divergentes. L'exemple emprunté à C. Notredame (Figure 13) illustre un tel cas de figure. Pour limiter ce type de problèmes, deux grandes catégories de stratégies ont été développées.

a. Concept de cohérence interne de l'alignement

La philosophie de la première stratégie, dite de « self-consistency », consiste à prévenir plutôt que guérir en optimisant dès le départ la cohérence entre les alignements initiaux. Pour intégrer cette notion de cohérence, une correction peut être introduite dans les scores par l'ajout de contributions qui vont dans l'exemple précédent (Figure 13) défavoriser la solution trouvée lors de l'alignement (A,B) et favoriser la solution plus fréquente associée aux alignements des groupes (A,C,D). Deux algorithmes d'alignements multiples parmi les plus précis, T-Coffee et ProbCons , utilisent ce type d'heuristique. T-Coffee (Tableau 3) est le programme qui a popularisée cette approche de « self-consistency ». L'algorithme permet d'intégrer de façon heuristique des sources d'alignement hétérogènes. Dans sa version originale, T-Coffee combine des alignements globaux obtenus par ClustalW avec des alignements locaux obtenus par DiAlign et prend en compte ses différents alignements pour évaluer favorablement ceux qui présentent le plus de cohérence entre eux. Plus récemment, l'algorithme ProbCons s'est inspiré du concept développé dans T-Coffee en se basant sur un formalisme HMM complet. Cet algorithme calcule des scores correctifs permettant d'intégrer l'information de cohérence entre les alignements par une approche statistique particulièrement élégante. Suite à l'alignement des séquences par paires avec un HMM, ProbCons exploite le formalisme HMM pour calculer les « probabilités postérieures » qui décrive la probabilité qu'un résidu i de la séquence A soit aligné avec un résidu j de la séquence B (noté P(Ai↔Bj| A,B)) dans l'alignement optimal. La question de la cohérence peut alors être traitée avec un formalisme Bayésien qui prend en compte les probabilités conditionnelles telles que pour une troisième séquence C il soit possible d'approximer :

$$P(Ai \leftrightarrow Bj | A,B,C) \approx \sum_k \left(P(Ai \leftrightarrow Ck | A,C) \cdot P(Bj \leftrightarrow Ck | B,C) \right)$$

Les probabilités postérieures sont également utilisées pour générer l'arbre guide de l'alignement multiple en alignant en priorité les séquences pour lesquelles le maximum de précision est attendu (celle pour lequel on doute le moins de les avoir aligné de façon cohérente). Ce maximum de précision attendu peut être évalué à partir de la somme des probabilités postérieures. La précision accrue obtenue grâce à l'intégration des contraintes de cohérence interne a un prix en terme de temps de calcul, les méthodes appliquant ce principe étant parmi les plus lentes (Figure 15).

b. Approches itératives

Dans une seconde stratégie visant à pallier aux erreurs d'alignement illustrées en Figure 13, ces erreurs sont tolérées lors des premières itérations et l'on tentera de les corriger par la suite. Typiquement, un sous groupe de séquences est sélectionné de façon aléatoire et réaligné (Figure 12). Si le score de la fonction objective augmente, l'alignement résultant est sélectionné. Cette procédure peut être répétée jusqu'à ce que le score global converge. Sur ce principe, l'algorithme MUSCLE récemment développé par Edgar , a atteint une rapidité remarquable en maintenant une qualité d'alignement légèrement inférieure aux meilleurs programmes. Le programme fonctionne en trois étapes restreignant au fur et à mesure l'espace de recherche et intégrant des étapes de raffinement de façon progressive. La première étape consiste, par exemple, à construire une ébauche de l'alignement multiple en privilégiant la rapidité sur la précision. Les séquences sont converties en un alphabet réduit qui permet de très rapidement identifier des segments similaires (kmer) et de construire l'arbre guide de l'alignement en un minimum de temps. Les séquences sont ensuite alignées entre elles de façon progressive. Enfin, les dernières étapes d'optimisation de MUSCLE, réalignent les séquences entre elles en extrayant de façon itérative deux sous alignements multiples et en les réalignant par des méthodes plus longues mais plus précises de comparaison profile-profile (méthodes décrites plus loin). Le principe de MUSCLE qui consiste à intégrer la complexité et le raffinement de façon hiérarchique est intéressant conceptuellement en tant que démarche prédictive en bioinformatique.

Mentionnons également l'algorithme MAFFT qui fonctionne sur une philosophie similaire à celle suivie dans MUSCLE . La première phase d'alignement rapide est exécutée par une analyse par transformée de Fourier qui recherche les corrélations entre les fréquences d'occurrence des différents acides aminés au sein de deux séquences. Comme dans MUSCLE, une représentation simplifiée des séquences est utilisée (chaque acide aminé est représenté par un vecteur à deux dimensions correspondant au volume et à la polarité). Des phases d'alignement progressif et d'optimisation itérative sont ensuite appliquées. Dans les dernières versions du programme, de nouveaux termes de raffinement de la procédure d'alignement ont été implémentés et ont permis d'atteindre une précision comparable aux autres méthodes sans perdre trop en rapidité, caractéristique importante de MAFFT (Figure 14 et Figure 15).

(iii) Comparaison entre les différentes méthodes.

La Figure 14 et la Figure 15 présentent les performances des différentes méthodes présentées précédemment sur les trois bases de données mentionnées dans cette section. Les données ont été extraites de la publication décrivant l'algorithme ProbCons et illustre l'intérêt

des programmes tels que MUSCLE et MAFFT pour des applications requérant l'alignement d'un nombre important de séquences. Dans cette publication ProbCons obtient les meilleures performances bien que dans l'absolu les niveaux d'amélioration d'une méthode à l'autre reste relativement faibles. L'étude présentée sur la Figure 16 montre néanmoins que ce gain en précision reste relatif d'une publication à l'autre en fonction des critères et des bases de données utilisés.

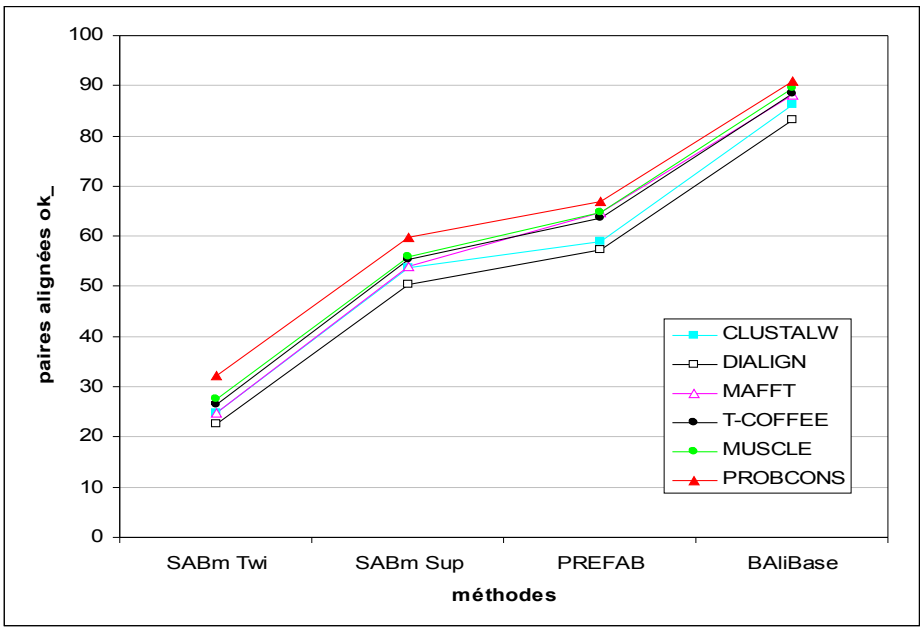


Figure 14 : Pourcentage moyen de résidus correctement alignés par rapport à l'ensemble des résidus alignés dans les blocs des alignements de référence calculé pour différentes bases de données tests SABmark (Twilight et Superfamily), PREFAB et BAlibase (extrait de).

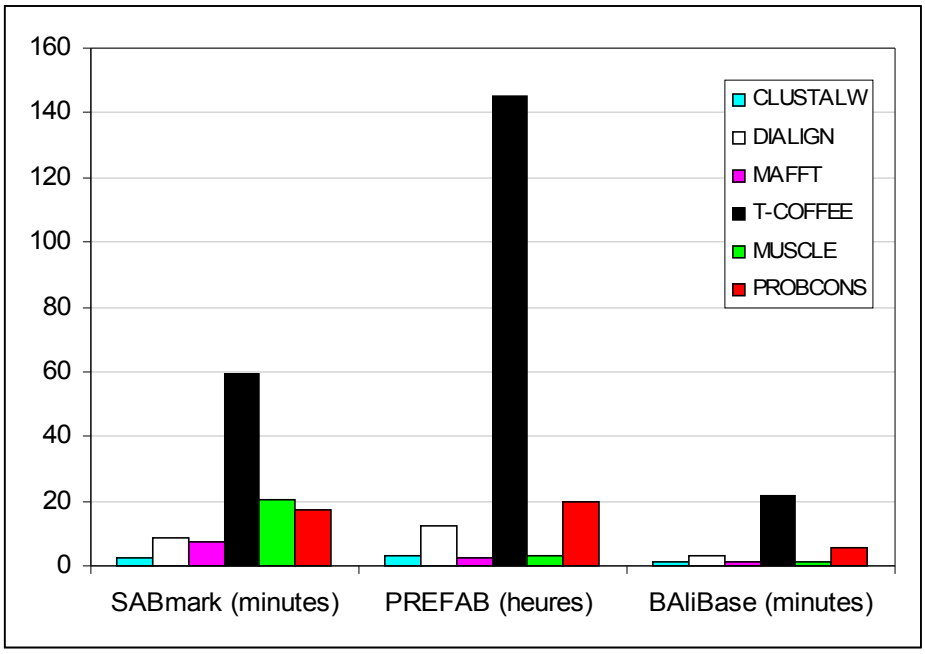


Figure 15 : Temps moyen d'exécution de l'opération d'alignement multiple pour les différents algorithmes utilisés en Figure 14 sur les différentes bases de données (extrait de).

I.1.1.8. Stratégies pour l'amélioration des alignements multiples à basse identité.

(i) Intégration d'informations structurales.

Une des tendances actuelles pour améliorer les alignements multiples repose sur l'intégration d'informations structurales pour guider les alignements. Une première approche consiste, quand un nombre suffisant de structures 3D sont connues, à intégrer deux ou plusieurs alignements structuraux dans l'alignement multiple. Le programme 3DCoffee directement développé sur la base de T-Coffee , peut, de par sa stratégie de « self-consistency », combiner et assembler des alignements par paires obtenus à partir de sources hétérogènes, comme des alignements structuraux ou des alignements séquence-structure (avec l'utilisation du programme de threading FUGUE . De façon paradoxale, le gain en précision obtenu par cette approche est relativement faible par rapport à ce qu'on aurait pu attendre, de l'ordre de 4 % par structure intégrée dans l'alignement.

Une autre stratégie d'intégration de l'information structurale repose sur les prédictions de structures secondaires. Celles-ci ont l'avantage de ne pas requérir de connaissance *a priori* de la structure tridimensionnelle. Un programme tel que SPEM développé par le groupe de Zhou (développeur du programme de threading SP3), intègre une stratégie très similaire à celle de T-Coffee mais module les pénalités des insertions en fonction du positionnement des

structures secondaires prédites. Il interdit par exemple les insertions au sein des structures secondaires. L'amélioration en pourcentage de résidus bien alignés par paires grâce à l'ajout des prédictions de structures secondaires varie entre 0.5 et 5.7% sur les différents ensembles de test de BAliBase. L'amélioration obtenue par la méthode SPEM est en fait principalement obtenue grâce à une étape initiale de comparaison profil-profil entre les profils associés à chacune des séquences à aligner.

On peut enfin noter le développement de la méthode MUMMALS , utilisant le formalisme HMM pour intégrer l'information de prédiction de structure secondaire pour effectuer l'alignement des paires de séquences. Ici encore, les gains obtenus par MUMMALS sont en moyenne de l'ordre de 3-4 % par rapport à ProbCons pour les alignements à basse identité.

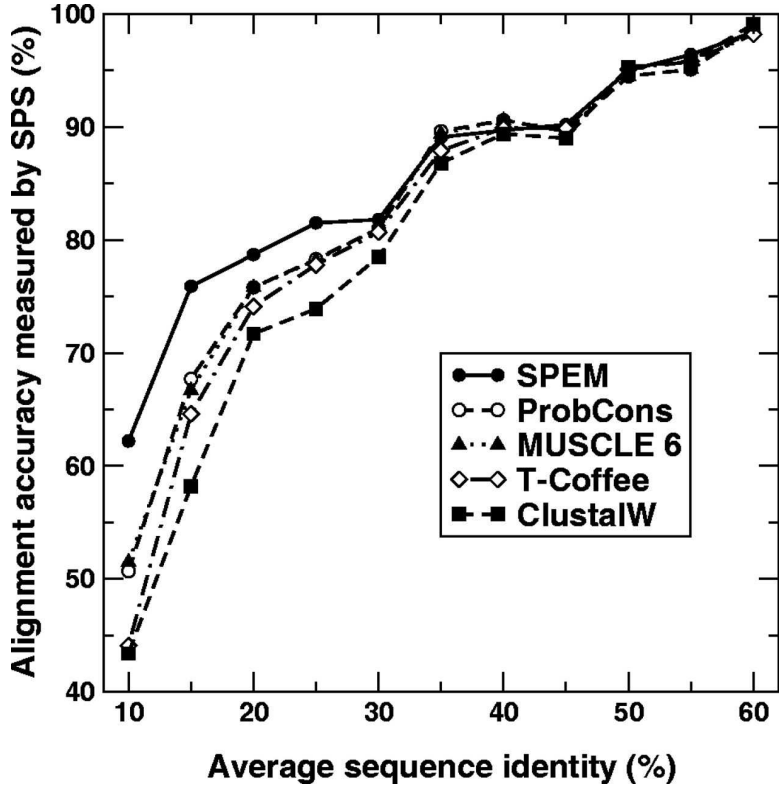


Figure 16 : Graphique présentant la précision des méthodes d'alignement multiple évaluée en fonction de l'identité moyenne entre les séquences de l'alignement (extrait de l'article de présentation de SPEM . Les alignements références sont issus de la base de données HOMSTRAD (233 familles > 3 séquences). Dans l'article les valeurs absolues de pourcentage de précision sont près de 10 % supérieures à celles publiées sur les mêmes bases tests et avec les mêmes programmes dans d'autres articles. Ceci est peut être dû à des différences dans le calcul du pourcentage d'accord entre les séquences. Néanmoins, la hiérarchie de précisions entre les méthodes est bien respectée.

(ii) La méta-prédiction

La méta-prédiction consiste à exploiter les solutions proposées par différentes méthodes et à les combiner afin d'atteindre un meilleur taux de succès que chaque méthode indépendante. Dans le domaine de la prédiction de structure, les méta-serveurs ont fréquemment remporté les premières places du classement des concours de prédiction de structure CASP . Dans le cas des alignements multiples, il a été observé que parmi les méthodes les plus performantes, les erreurs effectuées étaient souvent différentes. Le formalisme et la structure du programme T-Coffee sont particulièrement bien adaptés à l'extension de la méthode vers des stratégies de méta-prédiction. En effet, dès sa conception, le programme intégrait les résultats des algorithmes ClustalW et Dialign pour effectuer son analyse de cohérence interne. Dans un travail récent , une extension du programme T-Coffee, appelé M-Coffee, a montré qu'elle était à même d'augmenter de quelques pourcents supplémentaires les prédictions en intégrant les résultats de huit méthodes différentes. Cet intérêt pour la méta-prédiction est également souligné dans le programme MUMMALS qui intègre dans sa version serveur la possibilité d'effectuer ce type de méta-prédictions à partir de différents alignements. L'augmentation de la qualité des prédictions reste néanmoins modeste en regard du temps de calcul requis pour générer l'ensemble des données.

Pour conclure, de nombreuses approches algorithmiques ont été proposées pour accélérer et rendre plus fiables les alignements multiples. La tendance actuelle recherche de meilleures performances pour l'alignement des homologues lointains et intègre des informations nouvelles comme la structure tri-dimensionnelle des protéines ou la prédiction de structure secondaire. La méta-prédiction constitue également une piste intéressante pour dépasser les limites actuelles des méthodes d'alignements multiples. Dans cette thèse nous verrons que dans le domaine de la détection des homologues lointains nous avons également tiré profit de ces deux stratégies, prédictions de structures secondaires et méta-prédiction.

I.2.4. Développements des méthodes de comparaison profil-profil

Nous avons présenté les concepts importants de l'alignement multiple, montré quels sont les points d'accès aux différentes améliorations et situé l'évolution des derniers progrès

effectués dans le domaine. En particulier, nous avons vu que l'utilisation des profils a repoussé les limites des comparaisons entre séquences en augmentant considérablement la sensibilité de la détection d'homologies lointaines. Les nouvelles limites des capacités de détection des méthodes tels que PSI-BLAST ont stimulé l'essor d'une nouvelle catégorie de méthodes visant désormais à comparer l'information contenue entre deux profils. Nous allons maintenant aborder la généralisation de la notion d'alignement multiple dans le cadre du développement des méthodes de comparaison profil/profil.

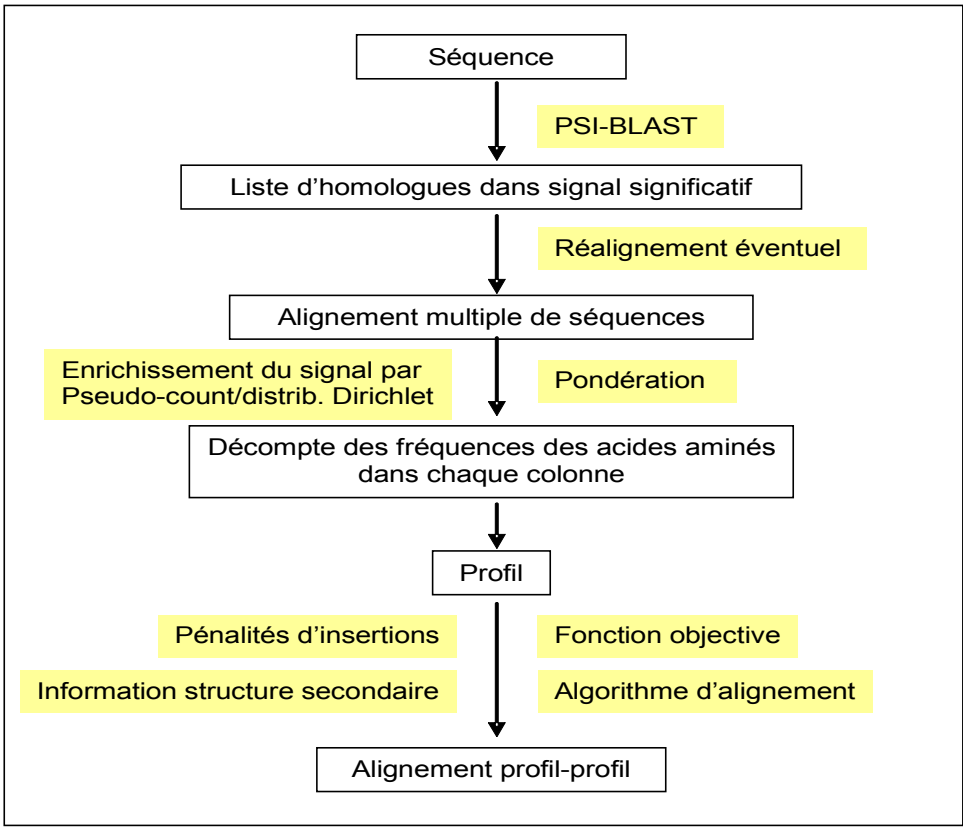


Figure 17 : Organigramme des étapes d'une comparaison profil-profil. Les cadres jaunes indiquent les composants qui distinguent les différentes méthodes entre elles.

Cette notion de comparaison profil-profil n'est pas nouvelle puisque dès le développement du programme Clustalw une étape de comparaison profil-profil était intégrée dans les stratégies d'alignements progressifs . Néanmoins, ces dernières années, de nouvelles méthodes de comparaison profil/profil ont été publiées présentant un fort potentiel pour la détection des homologies lointaines. Elles implémentent généralement des méthodes de scores plus performantes et, dans le meilleur des cas, intègrent un système d'évaluation statistique de

la validité des alignements profil/profil grâce à un paramètre de type e-value essentiel pour discriminer les solutions significatives. Un organigramme global décrivant les étapes des méthodes de comparaison profil/profil est présenté en Figure 17.

Les méthodes profil/profil au sens large permettent de comparer deux familles de séquences alignées. Ces familles peuvent être représentées soit par un profil, soit par un HMM. Les méthodes de comparaison entre deux alignements reposent sur l'une ou l'autre de ces représentations (Tableau 4). Je présenterai tout d'abord les méthodes basées sur une représentation de type profil. Puis j'évoquerai l'apport des méthodes utilisant des représentations de type HMM.

PROGRAMME	METHODE	FONCTION OBJECTIVE	DISPONIBLITE	REFERENCE
CLUSTALW	Prof-Prof	Moyenne des scores de toutes les combi. d'alignements par paires	LOCAL	
LAMA	Prof-Prof	Corrélation de Pearson entre fréquences	SERVEUR	
PROF_SIM	Prof-Prof	Probabiliste/Jensen-Shannon divergence entre distributions	LOCAL	
FFAS	Prof-Prof	Somme du produit des fréquences (DOT PRODUCT)	SERVEUR	
COMPASS	Prof-Prof	Probabiliste / + rapide que Jensen-shannon	LOCAL	
METABASIC	Prof-Prof	DOT PRODUCT	SERVEUR	
ORFEUS	Prof-Prof	Eq à FFAS + struct. secondaires	SERVEUR	
STRUCTFAST	Prof-Prof	Similaire à COMPASS	NON DISPO	
COACH	Prof-HMM	Moyenne probas d'emission de chq seq/HMM	LOCAL	
QC-COMP	Seq-HMM	Proba d'emission sur HMM2 de la seq consensus du HMM1	SERVEUR	
HHSEARCH	HMM-HMM	Probabiliste/Log de la somme des probabilités de co-occurrence.	LOCAL +SERVEUR	
PRC	HMM-HMM	Somme des probabilités de co-emission (produit des probas d'emission)	LOCAL	Pas de publi

Tableau 4 : Tableau récapitulatif des méthodes de comparaison profil-profil

I.1.1.9 Calcul de scores au sein des approches utilisant une représentation de type profil.

La plupart des méthodes de comparaison profil-profil sont basées sur la définition d'une fonction score mesurant la « distance » entre les colonnes de deux profils. La similitude entre deux profils correspond alors au score minimal obtenu par alignement des colonnes d'un profil sur les colonnes d'un autre profil. Cet alignement optimal peut être identifié par des techniques de programmation dynamique similaires à celle utilisée par Smith et Waterman pour l'alignement entre deux séquences. Cependant, il n'existe pas de mesure standard pour

évaluer le score correspondant à un alignement et plusieurs approches ont été proposées qui constituent l'une des distinctions principales entre les méthodes. Ces approches reposent soit sur un calcul de distances de type combinaison linéaire entre les paires de colonnes des profils, c'est-à-dire entre les paires de distributions de probabilités d'occurrence pour chaque acide aminé (ex : somme des paires, DOT PRODUCT), soit sur la corrélation des distributions de probabilité trouvées dans chaque paire de colonnes (cf Tableau 4).

Deux études indépendantes ont évalué les performances de ces approches sur leur propre ensemble de cas tests comportant environ 2000 alignements par paires . Bien que leurs systèmes d'évaluation soient distincts, les auteurs parviennent à des conclusions similaires. Les descriptions probabilistes utilisées par les programmes COMPASS et Prof_Sim pour évaluer la différence entre deux distributions de probabilités présentent de meilleures performances que les programmes utilisant des approches de type DOT PRODUCT. L'approche de COMPASS présente l'avantage d'être plus rapide à calculer que celle de Prof_Sim. Je me focaliserai donc dans la suite de ce sous-chapitre sur le programme COMPASS.

Hormis la fonction calculant la distance entre les colonnes des alignements, les principes utilisés dans COMPASS pour le calcul des scores, la prise en compte des insertions et le calcul de la e-value sont similaires à ceux développés dans PSI-BLAST. Par exemple, une généralisation directe des méthodes de « pseudo-counts » utilisées dans PSI-BLAST est appliquée aux deux colonnes à aligner entre elles. Le calcul analytique de la e-value à partir du score d'alignement dont les valeurs sont supposées suivre une distribution de la valeur extrême est également effectué à partir d'une estimation des paramètres K et λ avec la formule :

$$Evalue = K \cdot L_1 \cdot L_2 \cdot e^{-\lambda \cdot Score}$$

Comme dans PSI-BLAST, les paramètres K et λ varient en fonction de la longueur des profils L_1 et L_2 et les scores des matrices de substitutions ont été mis à l'échelle pour assurer une sensibilité optimale de la méthode . La validité du calcul de la e-value dans le cas de PSI-BLAST comme critère de quantification absolu du nombre attendu de faux-positifs a été remise en question. Par exemple, J. Soeding (développeur du programme HHsearch) a noté qu'en utilisant PSI-BLAST avec une procédure standard (jusqu'à 8 itérations, valeur seuil d'intégration de 10^{-4}) sur un grand nombre de séquences, il obtenait près de 100 fois trop d'homologues pour une e-value de 10^{-4} (http://toolkit.tuebingen.mpg.de/hhpred/help_ov). Une des raisons de ce biais provient de l'effet délétère de l'intégration d'une séquence non

homologue au sein du profil à l'une des étapes de la procédure itérative. Dans le chapitre IV de cette thèse, nous verrons que les e-value calculées par COMPASS et par un programme tel que HHsearch peuvent également poser question quant à l'interprétation statistique des seuils de e-value à utiliser pour filtrer les faux-positifs.

I.1.1.10 Les approches reposant sur une représentation de type HMM.

L'autre catégorie de méthodes de comparaison profil-profil plus récemment développée est basée sur le formalisme des chaînes de Markov cachées. Nous l'avons vu, la représentation des profils de familles protéiques par les profils HMM correspond à une description probabiliste très performante pour traiter les alignements multiples. Plusieurs méthodes de comparaison profil/profil utilisant le formalisme HMM ont été publiées au cours des dernières années. La plupart de ces méthodes ne constituent pas une réelle comparaison entre deux profils HMM et ramène le problème algorithmique à une comparaison profil/séquence en utilisant les mêmes modèles que ceux décrits précédemment. COACH par exemple est une méthode hybride qui compare un alignement multiple avec un HMM . Dans ce programme, les probabilités que le HMM déduit de l'alignement 2 puissent émettre chacune des séquences de l'alignement 1 de façon optimale sont calculées. Les probabilités extraites de ces alignements séquences-HMM sont ensuite combinées pour rendre compte des contraintes imposées par l'alignement 1. Une trajectoire globale d'alignement entre les deux alignements est alors déduite. Une autre approche implémentée dans le serveur QC-COMP, se ramène également à des comparaisons profil séquence . La séquence correspondant à un consensus du premier profil est alignée sur le second et vice versa. Seul le programme HHsearch intègre une véritable comparaison profil-profil associé à la construction d'un plan de transition/émission spécifique adapté à cette problématique. Enfin, PRC s'appuie également sur une réelle comparaison profil-profil mais n'a jusqu'à présent pas été publié et la documentation est seulement disponible sur internet (http://supfam.mrc-lmb.cam.ac.uk/PRC).

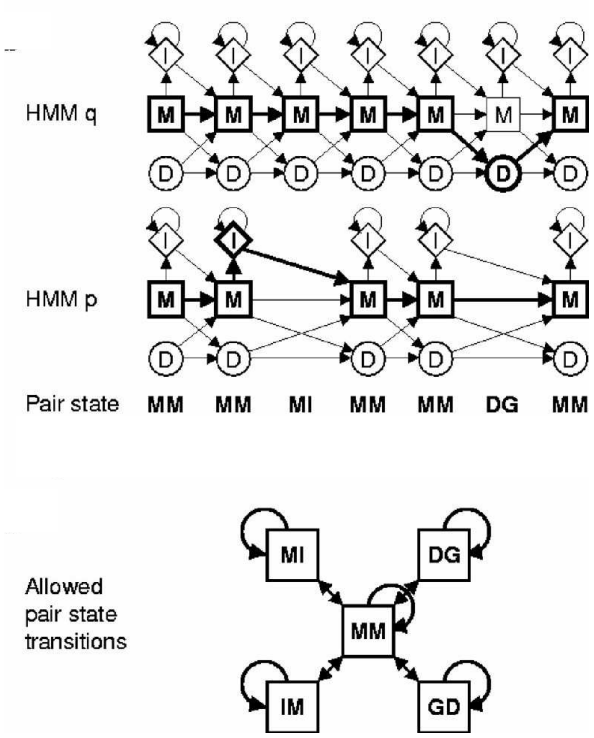


Figure 18 : Représentation des deux profils HMM q et p construits sur le modèle classique du plan 7. Les couples d'états autorisés correspondants à l'alignement des deux profils sont indiqués en dessous et constituent les états pris en compte dans HHsearch. Les transitions majeures autorisées entre ces couples d'état sont indiquées en bas. En utilisant la programmation dynamique, le programme HHsearch identifie la séquence d'états qui maximise la probabilité d'émission correspondant à l'alignement des deux profils.

Le plan de transition implémenté dans HHsearch (Figure 18) permet d'intégrer le concept de comparaison profil-profil dans la structure même du modèle de Markov caché. Ce programme a obtenu des résultats tout à fait remarquables au dernier concours de prédiction de structure automatique CAFASP5 comparé aux méthodes de threading qui intègrent des informations de structures tridimensionnelles de façon explicite (classement parmi les 5 meilleurs quelque soit la difficulté). HHsearch permet également d'intégrer les prédictions de structures secondaires sous forme de probabilité en tenant compte des niveaux de confiance avec lesquels les structures secondaires ont été prédites par un logiciel tel que PSI-PRED . L'intégration des structures secondaires repose sur un système assez sophistiqué puisque la probabilité qu'un résidu se trouve dans l'un des sept états de structure secondaire définit dans DSSP (H, E, C, G, B, S, T) est estimée à partir du niveau de confiance associé aux prédictions de PSI-PRED pour lequel seuls trois états sont prédits (H pour hélice, E pour les brins et C pour les structures désordonnées). Il n'est pas forcément aisé d'évaluer le gain en qualité de prédiction acquis grâce à ce degré de complexité. Néanmoins, l'ajout des

prédictions de structures secondaires améliore encore les performances du programme (Figure 19).

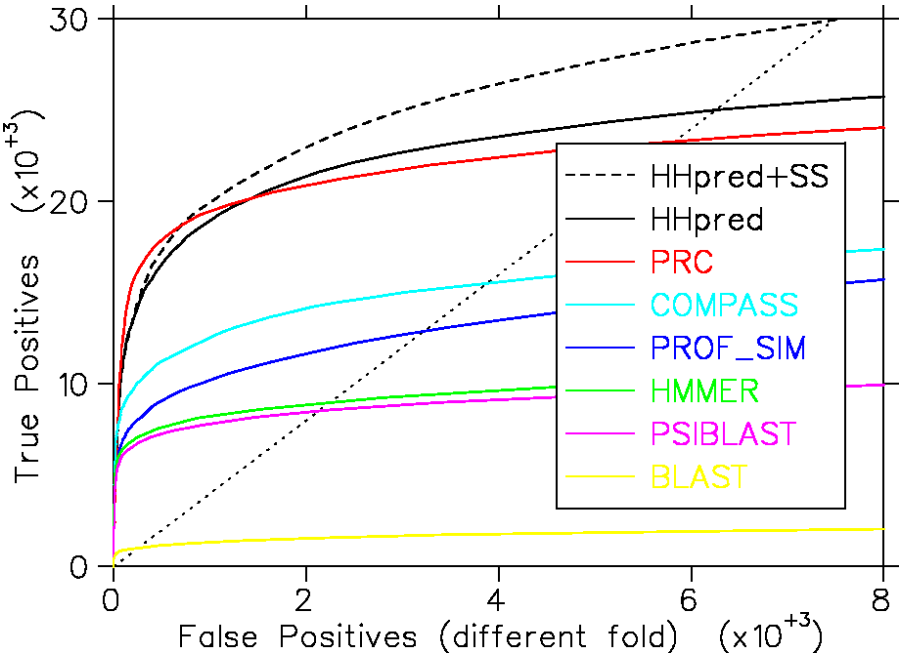


Figure 19 : Etude comparative des capacités de différentes méthodes de comparaison de séquences à discriminer des séquences possédant des repliements similaires. La base de données utilisée comprend 3691 domaines protéiques de la base de donnée SCOP dont les identités de séquences entre paires ne dépassent pas 20 %. Les courbes HHpred+SS et HHpred correspondent à la méthode HHsearch utilisant ou non les prédictions de structures secondaires, PRC est une méthode de comparaison HMM-HMM, COMPASS et PROF_SIM sont des méthodes de comparaison profil-profil, HMMER est une méthode de comparaison HMM-séquence, PSIBLAST, une méthode de comparaison profil-séquence et BLAST est une méthode de comparaison séquence-séquence.

I.1.1.11 Comparaison des différentes approches.

L'analyse comparée des performances des différents algorithmes d'alignements séquence-séquence, profil-séquence et profil-profil présentée en figure 18 a été publiée avec la description de la méthode HHsearch . Elle a été réalisée en générant toutes les combinaisons possibles d'alignements entre profils de séquences créés à partir d'une base de données de 3691 domaines extraits d'une version de la base de données SCOP filtrée à 20 % d'identité entre paires de séquences. Les résultats de cette analyse permettent clairement d'évaluer le gain en sensibilité apporté par les méthodes de comparaison profil/profil. Sur cet exemple, les programmes HHsearch et PRC (méthode non publiée disponible à l'adresse

suivante : http://supfam.org/PRC/) paraissent particulièrement performants. L'un des intérêts du programme HHsearch est également sa rapidité puisque la comparaison d'un profil avec les 3691 autres profils HMM est réalisée en 30s sur un processeur à 2 GHz (AMD64).

Dans cette thèse, j'ai cherché à tirer profit du développement des méthodes de comparaison entre alignements multiples pour développer un outil d'aide à la détection d'homologues lointains. J'ai sélectionné les deux méthodes COMPASS et HHsearch comme outils d'investigation, au vu de leurs performances et de leur complémentarité méthodologique. Ces programmes ont l'avantage supplémentaire d'être distribués en version locale et en serveur.

I.3. Utilisation des prédictions de structure 2D et 3D lors d'un alignement

La structure 3D des protéines étant mieux conservée que leur séquence , l'amélioration de l'alignement de séquences peu homologues peut être réalisée en incorporant des informations sur la structure 2D et 3D des différentes séquences. Nous l'avons vu dans le cas de HHsearch au chapitre précédent. De manière plus générale, des informations structurales ont été introduites dans les programmes d'alignement sous forme de structure secondaire dérivée ou prédite (, d'accessibilité au solvant dérivée ou prédite (Kelley et al., 2000 ; Karchin, 2003 #131} et plus récemment sous forme de probabilités de contact entre chaînes latérales dérivées de structures 3D de protéines . L'estimation de la compatibilité entre un alignement de séquences et d'autres alignements caractérisés par une structure 3D connue se fait par des algorithmes dits de « threading » ou enfilage de la séquence sur différents repliements successifs. A partir d'un échantillon de toutes les unités de repliements possibles, il faut enfiler successivement la séquence en cherchant à optimiser la répartition spatiale des acides aminés hydrophiles et hydrophobes tout en tenant compte du rapport surface/volume ainsi que du rayon de giration. La technique du « threading » ne retourne pas une solution mais un ensemble de structures candidates potentielles. Toutes ces méthodes ont montré leur efficacité dans la détection d'homologues lointains et la prédiction de la structure 3D lorsque la séquence d'intérêt ne présente pas d'homologie forte avec une structure 3D connue . Cependant, l'incorporation d'informations structurales par ces méthodes reste l'objet d'une discussion car la manière optimale de paramétrer la fonction score n'a pas été démontrée et dépend des données structurales à prendre en compte.

De plus, si la structure secondaire est aujourd'hui prédite avec une fiabilité proche de 70% sur séquence unique et qui peut atteindre 90% sur un alignement relativement divergent, il n'en est pas de même pour la structure tertiaire. L'expérience de CASP6 a montré que la difficulté de trouver un alignement entre une séquence et une structure 3D ne venait pas en général d'un problème d'échantillonnage des alignements, mais de la difficulté à discriminer l'alignement correct. En particulier, certains alignements corrects ne sont pas sélectionnés car la structure du squelette associée n'est localement pas compatible avec la séquence d'intérêt.

On peut discerner dans ce contexte l'un des avantages des méthodes d'alignement profil-profil pour la recherche d'homologues lointains : ces méthodes ne dépendent pas de la description des structures 3D, ni de la calibration des programmes de reconnaissance de repliement sur les structures 3D connues. De plus, elles ne requièrent pas la connaissance de la structure 3D des séquences homologues. Cet argument est d'autant plus important qu'il apparaît aujourd'hui que l'espace des repliements possibles n'est pas toujours discret, mais peut se présenter sous un certain continuum, et les différences observées entre des repliements, même proches, peuvent suffire à gêner la détection d'homologies lointaines par des techniques de « threading ».

L'ensemble des raisons mentionnées ici et les bonnes performances d'une méthode de comparaison profil/profil telle que HHsearch (équiv. HHpred) au concours de prédiction CAFASP5 en 2006 (http://www.cs.bgu.ac.il/~dfischer/CAFASP5/index.html) conforte l'hypothèse que la détection d'homologues lointains peut être effectuée en s'affranchissant de l'information structurale tertiaire explicite. En revanche, la prédiction de structure secondaire constitue un complément important à la comparaison profil/profil car la plupart des éléments de structures secondaires responsables de la formation d'un repliement se trouvent conservés au cours de l'évolution.

I.4. Objectifs de la thèse

Le travail présenté dans cette thèse est centré sur la détection d'homologues lointains, d'abord pour proposer des alignements et enrichir des profils de séquences, puis pour rechercher des protéines de structures connues reliées à la séquence d'intérêt. Lorsqu'aucune structure tridimensionnelle ne peut être associée à l'alignement détecté ou au profil généré, le but est d'identifier de nouveaux domaines structuraux dont l'étude expérimentale par RMN ou cristallographie permettra d'établir s'il existe réellement des relations structurales et

fonctionnelles avec d'autres protéines connues. Notre outil doit enfin pouvoir être utilisable à grande échelle afin d'obtenir un processus rapide d'identification de cibles d'intérêt pour les études structurales.

Plusieurs équipes de « chasseurs de domaines » ont proposé de rechercher les homologues lointains dans les alignements dits « non significatifs » calculés par le logiciel PSI-BLAST. Cet ensemble d'alignements constituant « le signal non significatif » regroupe l'ensemble des séquences pour lesquelles on ne peut pas conclure, sur un plan statistique, à l'existence d'une relation d'homologie avec la séquence étudiée. La détection de nouveaux homologues passe alors par une analyse approfondie de ces résultats non significatifs.

En s'inspirant de cette approche, nous avons cherché à développer une procédure automatique permettant de faciliter la détection d'homologues lointains parmi le signal non significatif de PSI-BLAST. Pour cela, nous avons tout d'abord réalisé une étude préliminaire visant à analyser la composition de ce signal non significatif et ainsi à mettre en valeur l'intérêt d'une approche permettant un filtrage efficace de ce signal (Chapitre II). Ensuite, nous avons exploré les potentialités de plusieurs étapes de filtrages successives du signal non significatif. Dans un premier temps, nous avons étudié si les prédictions de structures secondaires calculées pour la séquence d'intérêt et pour les séquences du signal non significatif pourraient permettre de discriminer les homologues lointains des autres séquences. Les résultats de cette analyse sont présentés au chapitre III. Par la suite, nous avons étudié les potentialités de deux méthodes d'alignement profil/profil COMPASS et HHsearch pour filtrer efficacement les séquences d'homologues lointains. Cette dernière étape de filtrage est présentée au chapitre IV.

Enfin, pour illustrer l'intérêt du programme développé, une application pratique basée sur l'analyse de 100 protéines de la signalisation des dommages de l'ADN a été entreprise. Nous présenterons les détails des résultats obtenus pour certaines de ces cibles au chapitre V. Cette étude de cas pratiques est essentielle pour montrer les potentialités de notre approche, pour identifier les limites actuelles auxquelles nous sommes confrontés et pour proposer des perspectives d'amélioration à court et moyen termes.

Chapitre II :Etude des alignements non-significatifs produits par le logiciel PSI-BLAST

II.1.Introduction

L'une des stratégies les plus efficaces pour l'identification d'homologues lointains consiste à analyser manuellement les séquences détectées de manière non significative par le logiciel PSI-Blast. En effet, le seuil de détection du logiciel PSI-BLAST doit être maintenu assez élevé pour éviter, lors de la procédure itérative, l'inclusion de séquences non homologues susceptibles de biaiser le profil. L'une des hypothèses sur laquelle repose cette analyse est alors que, pour des séquences ayant rapidement divergé au cours de l'évolution, le score de l'alignement obtenu peut être trop faible pour être détecté de façon significative.

Nous avons choisi de valider cette stratégie de façon systématique et automatisée, en réalisant une étude du signal non significatif obtenu avec le logiciel PSI-BLAST. Lors de cette étude, nous avons tenté d'identifier et de compter les homologues lointains présents dans les sorties de PSI-BLAST, en nous appuyant sur la notion de superfamille développée dans l'introduction. La base de données de domaines structuraux SCOP est particulièrement intéressante puisqu'elle intègre un niveau hiérarchique « Superfamily » reflétant les caractéristiques des homologues lointains : structures analogues, fonctions similaires, faibles identités de séquence. Dans ce chapitre, nous présentons les résultats de notre recherche d'homologues trouvés par le logiciel PSI-BLAST appliquée aux séquences de la base de données SCOP.

Une analyse de la composition du signal non significatif est tout d'abord présentée afin d'évaluer l'importance relative des séquences d'homologues lointains (appartenant à la même superfamille) et des séquences non homologues. Un point important de cette étude est non seulement d'analyser la présence de ces homologues lointains mais également d'évaluer la confiance qui peut être accordée aux alignements calculés par le logiciel PSI-BLAST lorsque les scores d'alignement sont faibles. Cette analyse a pour objectif final d'identifier un sous-ensemble de séquences présentant un nombre suffisant d'homologues lointains dans le signal non significatif. Ce sous-ensemble constituera la base de données de référence que nous utiliserons pour le développement d'une stratégie de détection automatique des homologues lointains dans les chapitres suivants.

II.2.Méthodes

II.2.1.Bases de données utilisées pour l'étude et protocole appliqué

Comme nous l'avons vu au chapitre 1, la version 1.69 de la banque SCOP (Structural Classification Of Protein) classe les 26000 structures de la PDB (Octobre 2004) en 7 classes, 945 repliements, 1539 superfamilles et 2845 familles. Le niveau hiérarchique des superfamilles nous intéresse tout particulièrement. En effet, pour que deux domaines appartiennent à la même superfamille, il faut que les structures et fonctions associées à ces domaines soient proches. Cependant, les séquences de ces domaines peuvent être très peu identiques (identité inférieure à 30% pour deux domaines n'appartenant pas à la même famille). Il existe donc un recoupement entre la définition d'une superfamille et la notion d'homologues lointains que nous souhaitons détecter. Dans la suite, nous utiliserons indifféremment les termes d'homologues lointains ou de membres de la même superfamille pour désigner ces cas de figure.

La base de données ASTRAL comprend l'ensemble des séquences protéiques associées aux domaines structuraux de la banque SCOP. L'en-tête des séquences contient le code de classification SCOP permettant d'identifier l'appartenance à un repliement ou à une superfamille donnée. Ainsi en travaillant à partir des séquences issues de cette base, il est aisé de déterminer si deux séquences appartiennent à une même superfamille et ainsi de conclure sur leur éventuelle relation d'homologie lointaine.

Pour éviter des analyses redondantes, nous avons choisi de travailler sur un sous-ensemble de séquences de la SCOP très différentes les unes des autres. Pour cela, nous avons utilisé la banque de séquences de la SCOP10 disponible sur ASTRAL : 3436 séquences filtrées de manière à ce que deux séquences ne présentent jamais une identité supérieure à 10%. La recherche d'homologues lointains a été réalisée sur une banque plus large, la SCOP40, également disponible sur ASTRAL. Nous avons choisi d'utiliser une banque filtrée à 40% d'identité afin d'identifier dans le signal non significatif du logiciel PSI-BLAST des homologues lointains qui soient assez divergents les uns des autres.

Quatre itérations du logiciel PSI-BLAST ont été effectuées avec une e-value d'inclusion de 0.001. Le signal non significatif rassemble tous les alignements possédant une e-value comprise entre 0.001 et 1000. Deux cents séquences présentant plus de 5 homologues lointains dans le signal non significatif du logiciel PSI-BLAST ont été sélectionnées pour la suite de l'analyse.

II.2.2.Evaluation de la qualité des alignements

(i) Utilisation d'une base d'alignements structuraux de référence.

L'évaluation de la qualité d'un alignement entre deux séquences de la base de données ASTRAL nécessite l'existence d'un alignement de référence. Les alignements de référence correspondent généralement aux alignements structuraux obtenus par superposition des structures 3D des protéines. Dans notre étude, ils ont été obtenus à partir de la base de données S4 qui contient les alignements structuraux des domaines appartenant à la même superfamille de la banque SCOP40.

Toutefois, la base S4 utilise des séquences extraites des fichiers de la PDB parfois différentes des séquences natives rassemblées au sein de la base de données ASTRAL. En effet des disparités peuvent exister entre ces sources de séquences. A titre d'exemple, l'absence de certaines régions désordonnées de la protéine ou encore la présence de mutations ponctuelles ou de résidus non naturels dans les fichiers PDB. Ces différences ne permettent pas d'utiliser directement la banque S4 pour évaluer la qualité des alignements obtenus avec les séquences issues d'ASTRAL. Nous avons donc modifié les données de la base S4 de manière à conserver les alignements tout en rectifiant les séquences qui n'étaient pas identiques aux séquences de la base ASTRAL. Pour cela nous avons réalisé trois types de modifications :

- mutation de certains résidus de la base S4 pour que ces résidus soient toujours identiques à ceux trouvés dans la base ASTRAL ;
- remplacement des insertions présentes dans les séquences de la base S4 par des gaps ;
- ajout des résidus manquants dans la base S4 mais présents dans la base ASTRAL. Chaque insertion de résidus entraîne l'ajout d'insertions au niveau des autres séquences constituant l'alignement.

(ii) Méthode de Scoring.

Afin d'évaluer la qualité des alignements, nous avons calculé deux types de scores couramment utilisés dans la littérature : le Qdev et le Qmod. Le Qdev (pour Qdeveloper) permet d'évaluer la qualité de l'alignement de façon globale alors que le Qmod (pour Qmodeler) permet une évaluation au niveau local (Figure 20). L'utilisation de ces deux systèmes de scores est complémentaire.

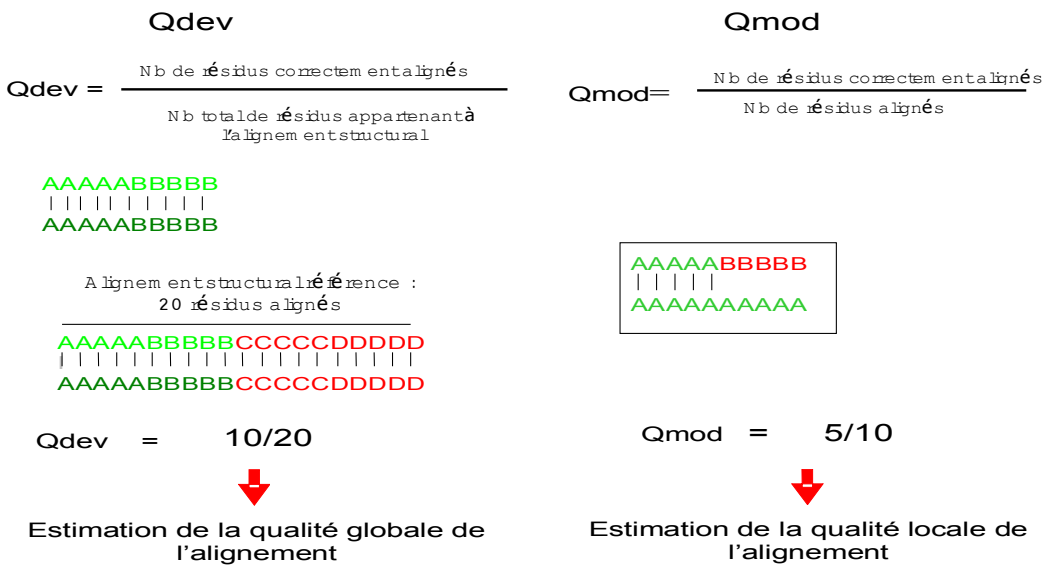


Figure 20: Calcul des Qdev et Qmod. Pour déterminer le Qdev, le nombre de résidus alignés de manière identique à l'alignement structural de référence est calculé, ce nombre est ensuite divisé par la longueur totale de la séquence alignée. Pour déterminer le Qmod, le nombre de résidus alignés de manière identique à l'alignement structural de référence est calculé, ce nombre est ensuite divisé par la longueur de l'alignement proposé par la méthode d'alignement employée.

II.3. Résultats : analyse préliminaire du signal non significatif obtenu avec le logiciel PSI-BLAST

II.3.1. Quantification du nombre d'homologues lointains situés dans le signal non significatif du logiciel PSI-BLAST

Pour chaque séquence de la SCOP10, une recherche d'homologues a été effectuée sur la SCOP40 à l'aide du logiciel PSI-BLAST. Puis, l'ensemble des séquences présentes dans le signal non significatif (e-values de 0.001 à 1000) a été analysé. Le nombre de séquences retrouvées dans le signal non significatif et appartenant à la même superfamille que la séquence de la SCOP10 soumise à PSI-BLAST a été calculé. La distribution du nombre de séquences ainsi identifiées lors de l'analyse de toutes les séquences de la SCOP10 est présentée en Figure 21.

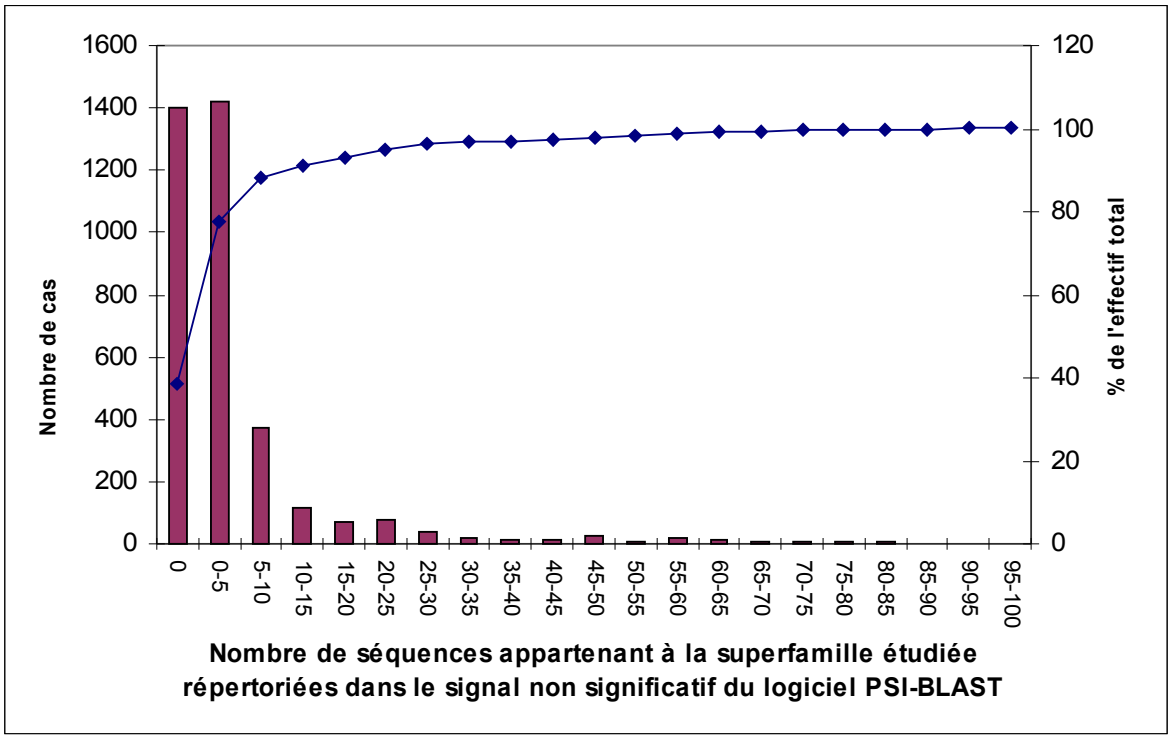


Figure 21 : Distribution (histogramme bordeaux) du nombre de séquences (ordonnée gauche) de la même superfamille répertoriées dans le signal non significatif obtenu avec le logiciel PSI-BLAST. En abscisse figure le nombre de séquences de la même superfamille que la séquence d'intérêt identifiées dans le signal non significatif. La courbe bleue représente le pourcentage de séquences de la SCOP10 possédant au plus un intervalle donné de nombres d'homologues lointains dans le signal non significatif (noté sur l'ordonnée à droite).

La Figure 21 indique que 65% des séquences de la SCOP10, présentent au moins un membre de leur superfamille dans le signal non significatif. La distribution du nombre de séquences montre que pour près de 1800 cas, il existe entre 1 et 10 membres de la même superfamille dans le signal non significatif. A l'autre extrémité de la distribution, nous observons que dans quelques cas exceptionnels, jusqu'à 90 membres de la même superfamille peuvent être retrouvés au sein du signal non significatif. Ces cas extrêmes sont associés à quelques superfamilles de grande taille contenant des familles très divergentes en termes de séquences. C'est le cas notamment de la séquence d1e9a appartenant à la superfamille c.37.1 des hydrolases de nucléosides triphosphates présentant une « P-LOOP » dont 10 membres de la superfamille sont détectés de manière significative et 86 de manière non significative.

Une analyse complémentaire sur les séquences de la SCOP10 présentant plus de 5 séquences de leur superfamille dans le signal non significatif a été réalisée. Celle-ci nous a permis d'observer que le rapport entre le nombre de séquences homologues présentes d'une part dans le signal significatif et d'autre part dans le signal non significatif est déséquilibré : en moyenne 3 fois plus de séquences de leur superfamille appartiennent au signal non significatif.

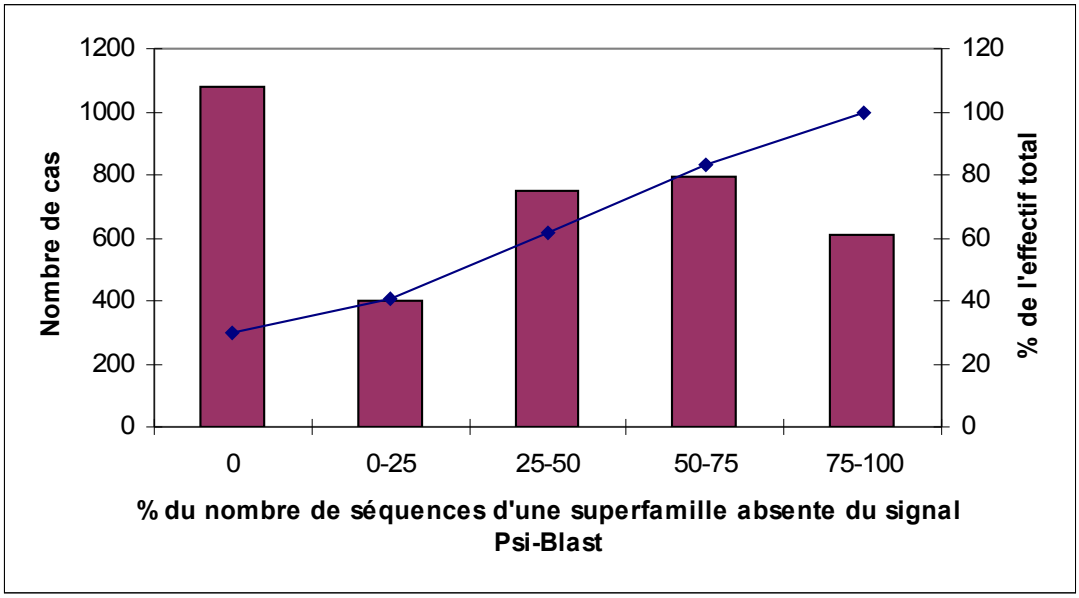


Figure 22 : Distribution des membres de la même superfamille non détectés par PSI-BLAST (avec un seuil de e-value maximal fixé à 1000). Ces données ont été calculées pour les 3436 séquences de la banque SCOP filtrée à 10% d'identité. La courbe bleue représente le pourcentage des effectifs cumulés. En abscisse figurent les pourcentages de séquences non détectées par rapport à l'ensemble des séquences de la superfamille. En ordonnée figurent, à gauche le % cumulé de l'effectif, à droite le pourcentage de ces séquences par rapport à l'effectif total.

Afin de voir si l'utilisation de la banque SCOP et du logiciel PSI-BLAST est appropriée pour notre étude, nous avons déterminé le nombre de séquences de la superfamille détectées (e-value maximale de 1000) par rapport au nombre total de séquences constituant la superfamille dans la banque SCOP40. Sur le graphique de la , nous pouvons ainsi constater que, dans environ 35 % des cas, l'ensemble des séquences appartenant à la superfamille est détecté par le logiciel PSI-BLAST. Nous remarquons également que pour un nombre conséquent de séquences (près de 20 % de l'effectif), plus de 75 % des membres de la superfamille ne sont pas du tout détectés par le programme PSI-BLAST. Dans le cas de séquences présentant de fortes divergences, il n'est donc pas assuré que le signal non significatif contienne de façon exhaustive les séquences potentiellement intéressantes.

Nous avons ensuite évalué l'intérêt de travailler sur le signal non significatif produit par PSI-BLAST plutôt que sur un échantillon aléatoire de la base de données de séquences SCOP. En d'autres termes, nous avons cherché à évaluer si le signal non significatif concentrait effectivement plus d'homologues lointains que la base de données elle-même. Dans notre étude, sur environ 17526 séquences de la base de données SCOP100, le bruit de fond sélectionné comprend déjà en moyenne 710 séquences (~600 séquences sur la SCOP40), et l'enrichissement en homologues lointains est d'environ un facteur 6.

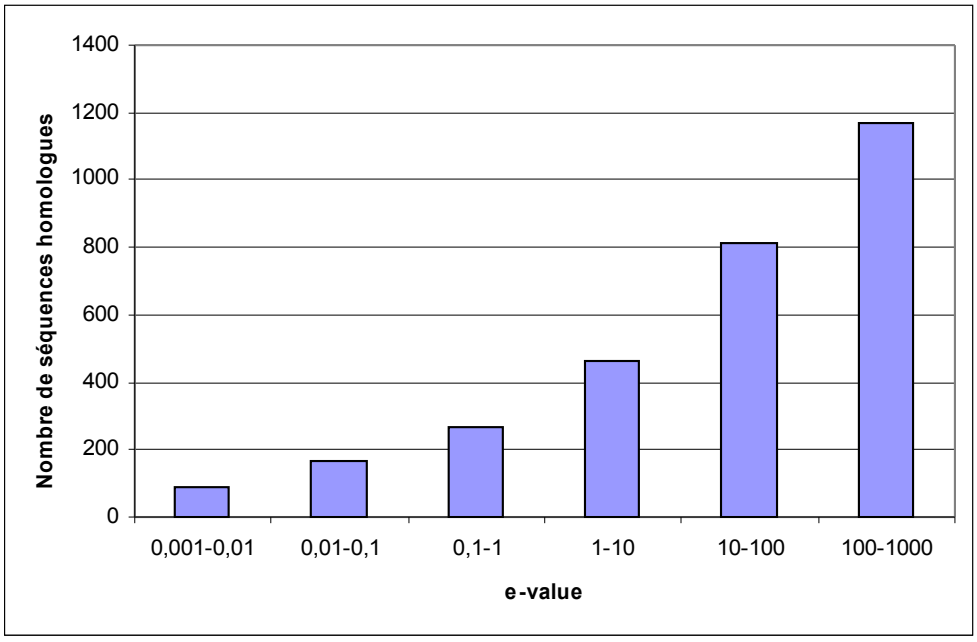


Figure 23 : Distribution du nombre des membres d'une même superfamille présentes dans la sortie PSI-BLAST (ordonnée) en fonction des e-values calculées par PSI-BLAST (abscisse) par intervalles de valeurs (de 0,001 à 1000 par pas d'un facteur 10).

La sélection des séquences dans le signal non significatif a été effectuée sur un critère de e-values comprises entre 10^{-3} et 1000. Etant donné le nombre très important de données composant ce signal, nous avons évalué l'intérêt de réduire la valeur seuil 1000 pour restreindre le nombre de séquences à analyser par la suite (Figure 23).

La distribution du nombre de membres d'une même superfamille en fonction des intervalles de e-values étudiés est croissante. D'après cette analyse, il semble intéressant de s'intéresser aux e-values comprises entre 100 et 1000 puisqu'elles peuvent contenir jusqu'à 40% des séquences d'homologues lointains détectés dans le signal non significatif. Toutefois ces séquences d'homologues lointains sont de plus en plus diluées à mesure que la e-value augmente. La Figure 24 permet de quantifier cet effet de dilution par l'analyse du pourcentage de séquences d'homologues lointains par rapport à l'ensemble des séquences identifiées à différentes e-value.

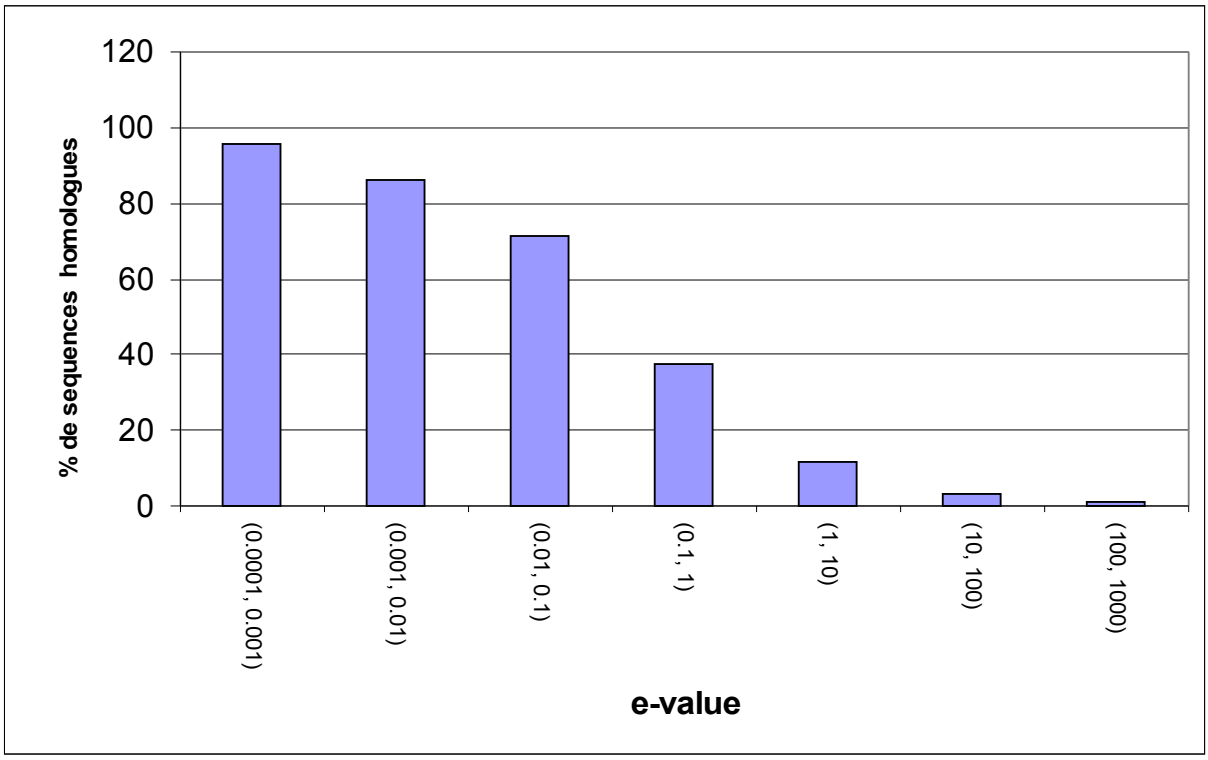


Figure 24 : Distribution du pourcentage de séquences homologues présentes dans la sortie PSI-BLAST en fonction d'un intervalle de e-value donné. Pour chaque intervalle de e-value (en abscisse), le nombre de séquences appartenant à la superfamille étudiée ainsi que le nombre de séquences totales sont calculés. Le pourcentage de séquences homologues est ensuite calculé (en ordonnée).

La distribution du pourcentage de séquences homologues en fonction de la e-value est décroissante. Pour l'intervalle de e-values significatives [0,0001 ; 0,001] on observe que le signal est constitué d'environ 98% de séquences de la même superfamille. Dans les intervalles

de e-values correspondants au signal non significatif, nous observons que dans l'intervalle [0,1 ;1], environ 40% de séquences en moyenne sont des homologues et que dans l'intervalle [100 ; 1000], environ 1 % des séquences sont des homologues. Ces valeurs sont en accord avec la définition statistique d'une e-value. Elles renforcent l'idée que la définition de superfamille établie dans SCOP est cohérente avec la notion d'homologie lointaine. Nous verrons dans le chapitre IV que la cohérence entre e-value et nombre de séquences détectées comme homologues lointains n'est pas toujours respectée dans les programmes de comparaison profil-profil.

II.3.2. Evaluation de la validité des alignements calculés par PSI-BLAST dans le signal non significatif.

Nous avons testé, dans le cas des 200 séquences définies précédemment, la cohérence entre l'alignement proposé par le logiciel PSI-BLAST pour deux séquences de la même superfamille et l'alignement structural correspondant. Pour cela, nous avons comparé la distribution des valeurs de Qmod et Qdev (définit précédemment page 64) obtenues pour les alignements entre membres d'une même superfamille dont le score est significatif (e-value comprise entre 0 et 10^{-3}) et ceux dont le score est non significatif (e-value comprise entre 10^{-3} et 1000). Rappelons que le calcul a été réalisé en comparant les alignements proposés par le logiciel PSI-BLAST aux alignements structuraux de la banque S4 modifiée (cf Méthodes).

I.1.1.12 Analyse de la qualité locale des alignements proposés par le logiciel PSI-BLAST

Dans un premier temps, nous avons déterminé les Qmod associés aux alignements du signal non-significatif afin de vérifier que les alignements proposés pour des séquences d'homologues lointains par le logiciel PSI-BLAST correspondent à une réalité structurale.

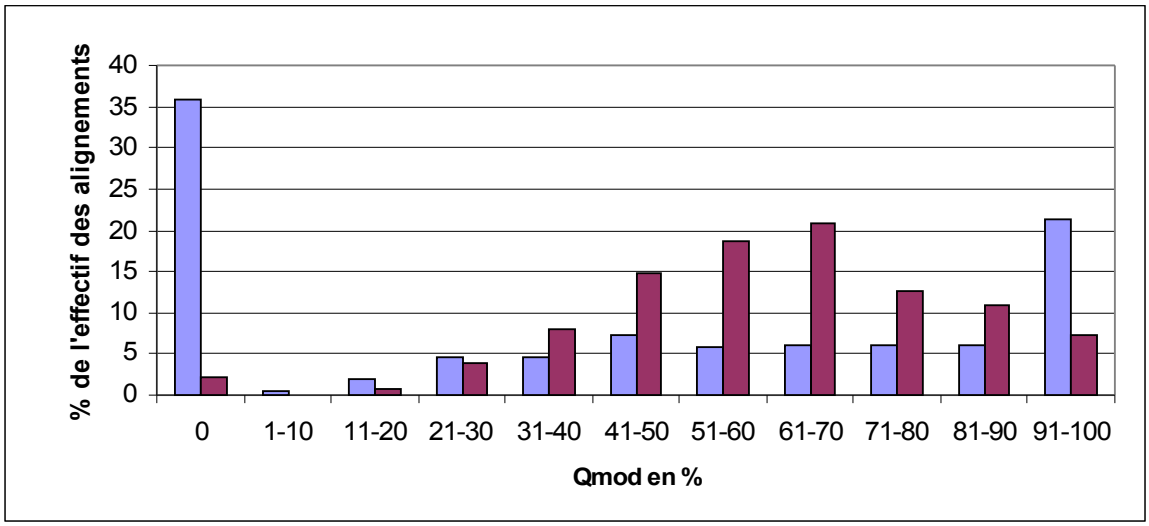


Figure 25 : Distribution des valeurs de Qmod pour les alignements non significatifs (e-value > 0,001) (bleu) et significatifs (e-value < 0,001) (rouge) calculés par le logiciel PSI-BLAST pour les 200 séquences étudiées. Les intervalles de valeurs de Qmod exprimées en % sont indiqués en abscisse et les pourcentages d'alignements correspondants à chaque intervalle de valeurs de Qmod sont présentés en ordonnée.

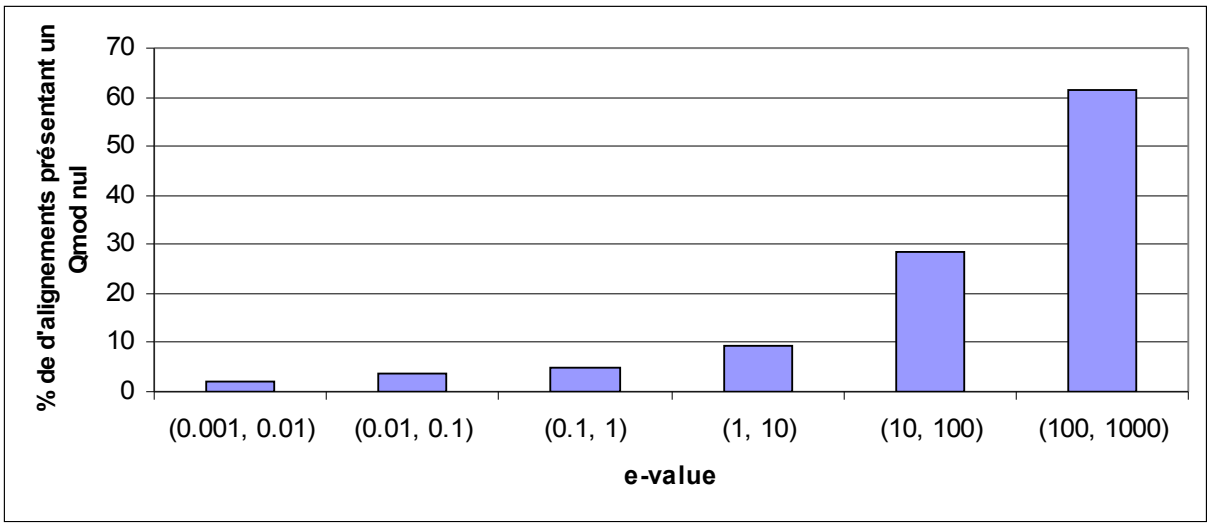


Figure 26 : Distribution du pourcentage d'alignements de séquences de la même superfamille présentant un Qmod nul en fonction de la e-value. Pour chaque intervalle de e-value le nombre d'alignements présentant un Qmod nul est dénombré puis rapporté au nombre total d'alignements présents dans l'intervalle de e-value étudiée.

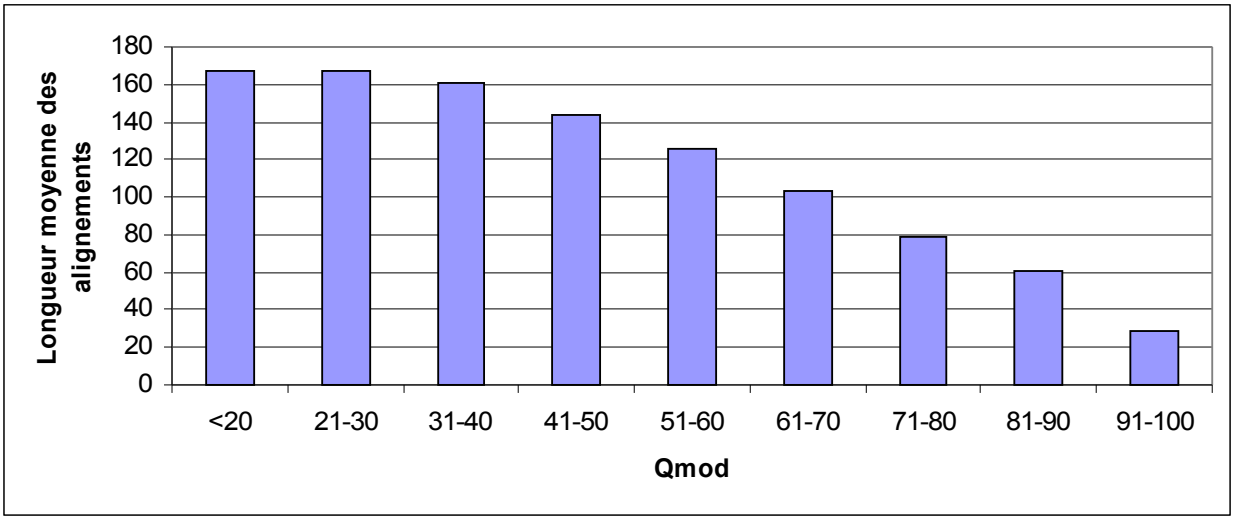


Figure 27 : Distribution de la longueur moyenne des alignements du signal non-significatif en fonction du Qmod. Pour calculer la longueur moyenne des alignements associés à un Qmod nul sont exclus. Les alignements présentant des Qmod <20 sont fusionnés en 1 classe unique.

La Figure 25 présente la distribution des Qmod associés aux séquences détectées de manière non significative (en bleu) et de manière significative (en rouge). La distribution des Qmod pour les séquences détectées de manière non significative comprend deux pics, l'un au niveau des valeurs nulles de Qmod (environ 1000 cas soit ~35% de l'effectif) et l'autre au niveau des valeurs de Qmod comprises entre 90 et 100 % (600 cas). Le nombre de cas reste relativement constant pour des valeurs de Qmod comprises entre 20 et 80 % (environ 150 cas soit près de 5% de l'effectif dans chaque intervalle). Le maximum observé pour les Qmod nuls indique que près d'un tiers des séquences identifiées par PSI-BLAST est en réalité

détecté de façon fortuite car l'alignement proposé ne correspond pas à la réalité de l'alignement structural. Néanmoins, pour deux tiers des séquences, les alignements de PSI-BLAST correspondent au moins en partie à l'alignement structural. La Figure 26 présentant le nombre d'alignements fortuits en fonction de la e-value montre que la proportion d'alignements fortuits suit une distribution croissante en fonction de la e-value. Un point étonnant concerne l'existence d'alignements fortuits considérés comme significatifs par PSI-BLAST. Une analyse spécifique de ces cas montre qu'il s'agit de repliements possédant des structures répétées et qu'un décalage global de l'alignement induit ces valeurs nulles. Nous pouvons aussi observer que pour un intervalle de e-value compris entre 0,001 et 10, seul un alignement sur cinq ne sera pas associé à une correspondance structurale. Pour des e-values supérieures à 10, le pourcentage d'alignements fortuits augmente rapidement, l'intervalle de e-value compris entre 100 et 1000 présentant en moyenne 2 alignements fortuits sur 3.

Ce résultat est important à prendre en compte pour évaluer par la suite la pertinence des stratégies de récupération des séquences d'homologues lointains. La Figure 27 représentant la longueur moyenne des alignements en fonction du Qmod montre que, parmi les valeurs de Qmod les plus élevées (comprises entre 90 et 100 %), on retrouve principalement des alignements de petite taille (27 résidus en moyenne). Cette observation suggère que ces alignements de petite taille correspondent à des HSP (cf introduction sur Blast, chapitre I.2.2.3) qui n'ont pas pu être étendus ou agrégés avec d'autres alignements. Dans l'intervalle de Qmod compris entre 20 et 90 % où chacun des intervalles représente environ 5% de l'effectif, la Figure 27 nous permet d'observer que les alignements présentent des tailles moyennes décroissantes de 160 à 60 résidus. Ces observations suggèrent que pour les alignements du signal non significatif présentant une taille supérieure à 100 résidus un minimum de 50% des résidus est aligné de manière erronée par rapport à une référence structurale. Nous pouvons en conclure que parmi le signal non significatif du logiciel PSI-BLAST les alignements de grande taille sont généralement de mauvaise qualité et qu'on ne doit pas négliger les alignements de petite taille car ils peuvent être associés sur toute leur longueur à une réalité structurale.

I.1.1.13Analyse des alignements proposés par le logiciel PSI-BLAST par rapport à la longueur de l'alignement structural

L'analyse des alignements par le paramètre de Qdev doit permettre d'évaluer quelle fraction de l'alignement structural global se retrouve bien alignée dans le signal non significatif.

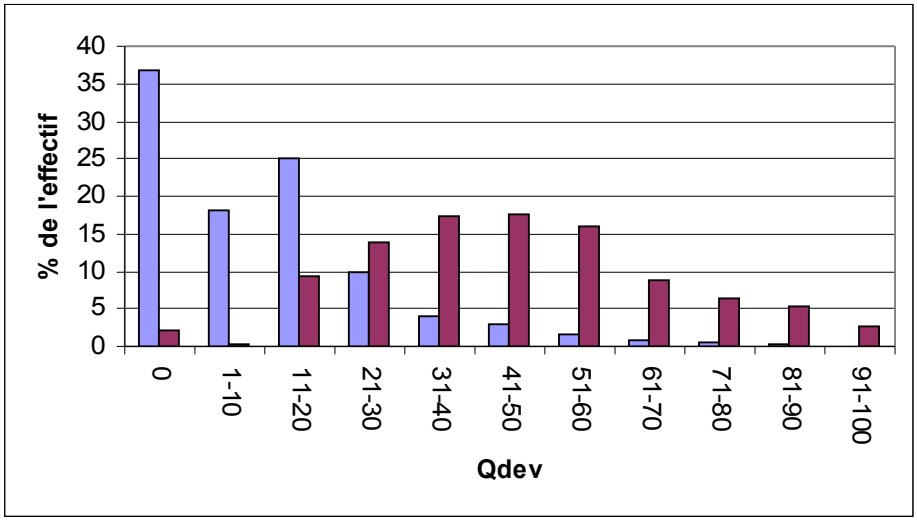


Figure 28: Distribution des Qdev des alignements considérés comme significatifs (e-value <0.001) (en rouge) et non significatifs (e-value > 0.001) (en bleu) par le logiciel PSI-BLAST. Les intervalles de valeurs de Qdev sont indiqués en abscisse et le pourcentage d'alignement dans chaque intervalle de e-value est présenté en ordonnée.

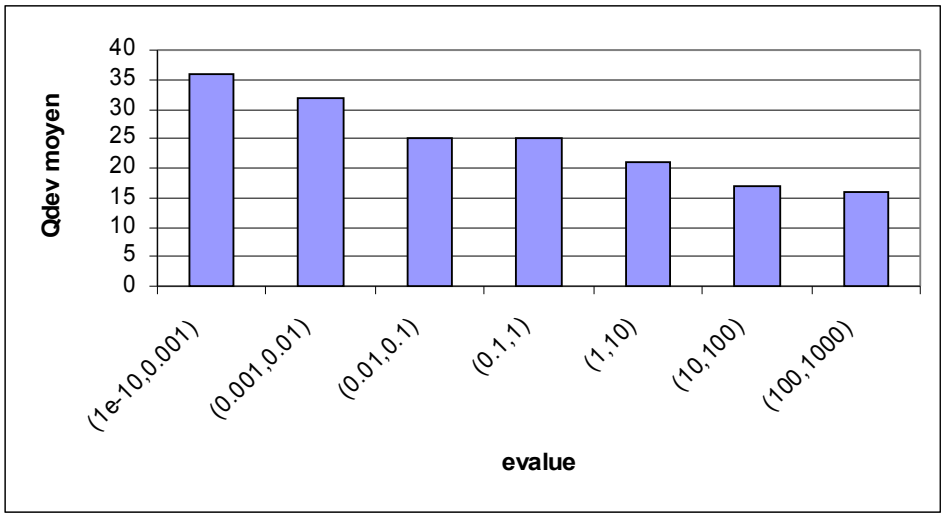


Figure 29 : Distribution de la valeur moyenne des Qdev associés aux séquences homologues détectées par le logiciel PSI-BLAST en fonction de la e-value. En ordonnée figurent les Qdev moyens, en abscisse les e-values. Le Qdev moyen est déterminé en excluant les alignements fortuits présentant un Qdev nul.

La distribution des valeurs de Qdev associées aux séquences du signal significatif (en rouge) et non significatif (en bleu) est présentée sur la Figure 28. En omettant le pic des alignements fortuits présentant un Qdev nul, le profil de la distribution des séquences du signal non-significatif est globalement décroissant et montre un pic correspondant aux alignements qui recouvrent de 11 à 20 % de l'alignement structural. Les séquences d'une même superfamille sont donc présentes dans le signal non significatif grâce à la détection d'une homologie avec une fraction faible des éléments de structure qui composent ces domaines. Cette situation contraste avec la distribution des valeurs de Qdev obtenues pour les séquences détectées dans le signal significatif.

Dans ce cas, l'alignement recouvre en moyenne près de 50 % de l'alignement structural. Pour un petit nombre d'alignements considérés comme significatifs par PSI-BLAST, le Qdev est nul. Ces alignements ont aussi un Qmod nul et correspondent à des repliements possédant des structures répétées. La Figure 29 montre la dépendance globale existant entre les valeurs moyennes de Qdev et les e-value associées aux alignements. Pour des e-values très élevées (au-delà de 100), le taux de recouvrement des régions correctement alignées devient très faible (de l'ordre de 15%). Autour des régions alignées, la divergence est donc probablement trop forte pour permettre à l'algorithme de Blast d'agréger d'autres HSP.

II.4. Exemple de deux homologues lointains : deux domaines de liaison au NADP : 1hxha et 1dih.

Pour illustrer les observations effectuées de façon statistique sur l'ensemble des 200 séquences, nous avons choisi de présenter un exemple de superfamille pour laquelle plusieurs séquences sont détectées dans le signal non significatif avec des valeurs de Qmod relativement élevées. Ce type d'analyse peut nous permettre de mieux appréhender quelles approches sont susceptibles d'aider à améliorer les limites des méthodes de détection.

Les domaines d1hxha et d1dih1 sont tous deux issus de la banque SCOP. Ces domaines appartiennent à la classe des protéines α et β (α/β) (Figure 30). Ils présentent un repliement s'organisant en 3 couches (α/β/α) et s'articulant autour d'un cœur composé de 6 brins β. Ils sont décrits comme appartenant à la superfamille C.2.1 qui regroupe les domaines de liaison au NADP présentant un repliement de type « Rossman-fold ». Le domaine 1hxha appartient à la famille des « oxydoréductases tyrosine dépendante ». De son côté, le domaine

d1dih appartient à la famille des domaines N-terminaux des protéines ressemblant aux « glycéraldéhyde-3-phosphate deshydrogènases ».

L'analyse des résultats obtenus après une recherche PSI-BLAST effectuée à partir du domaine 1hxha sur la SCOP40 a permis de détecter de manière significative 32 membres de la superfamille. 46 membres de la superfamille sont présents dans le signal non significatif (evalue comprise entre 10^{-3} et 1000) et 36 membres de la superfamille ne sont pas détectés pour des e-values supérieures à 1000. Le domaine 1dih se trouve précisément dans le signal non significatif de PSI-BLAST calculé à partir de la séquence du domaine 1hxha. Pour évaluer comment PSI-BLAST aligne entre eux ces deux domaines, intéressons nous à leur structure.

L'alignement structural des domaines 1dih et 1hxha montre que ces deux domaines possèdent en commun une région composée de 5 hélices et 5 brins (Figure 31). A part ce cœur conservé, il existe une grande variabilité au sein des boucles avec des insertions de grande taille incluant des domaines repliés de façon autonome (Figure 30). L'identité de séquence globale est de l'ordre de 10 %.

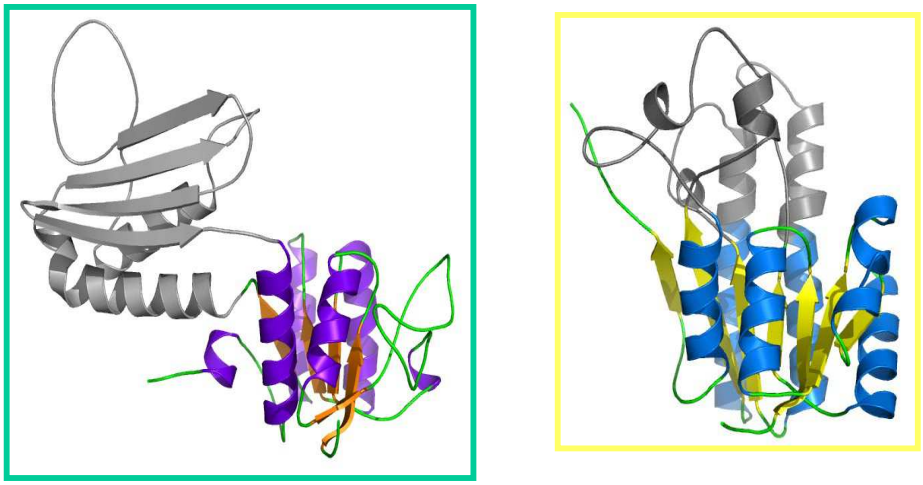


Figure 30 : La structure à gauche représente le domaine 1dih. En violet sont figurées les hélices alpha, en orange les brins béta, en vert les boucles. La structure à droite représente le domaine 1hxha. En bleu sont figurées les hélices alpha, en jaune les brins béta et en vert les boucles. En gris sont représentées les structures secondaires n'appartenant pas au « Rossman-fold.

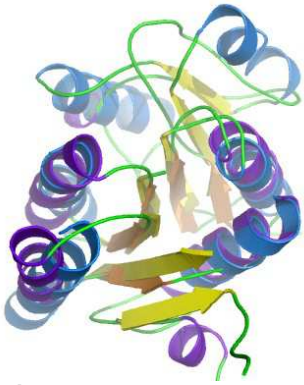


Figure 31 : Superposition de la région structurale conservée entre les domaines d1hxha et d1dih.

L'alignement non significatif calculé par le logiciel PSI-BLAST est présenté Figure 32. Le logiciel PSI-BLAST considère que l'alignement entre ces deux domaines est non significatif (e-value= 365) et propose un alignement de ces deux séquences sur une portion de 26 acides aminés présentant 18% d'identité.

> d1dih_1 c.2.1.3 Dihydrodipicolinate reductase [E. coli]

E-value = 365 Identities = (18%)

```
Query (1hxha): 8   VALVTGGASGVGLEVVK-LLLGEGAKV 33
                      + G     +G ++++   L  EG ++
Sbjct (d1dih): 6   RVAIAGAGGRMGRQLIQAALALEGVQL 32
```

Figure 32 : Descriptif de l'alignement et de la e-value obtenue pour la séquence de d1dih lors de la requête PSI-BLAST sur la banque SCOP40 effectuée à partir de 1hxha.

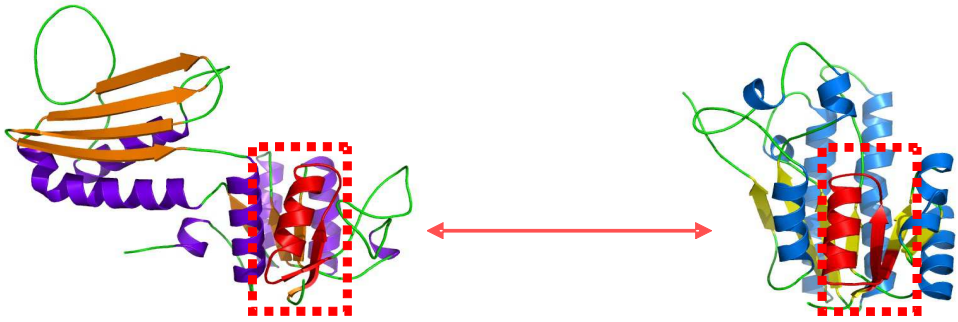


Figure 33 : Sur la gauche la structure du domaine 1dih. En violet sont figurées les hélices α, en orange les brins béta et en vert les boucles. Sur la droite la structure du domaine 1hxha. En bleu sont figurées les hélices α, en jaune les brins béta et en vert boucles. Dans l'encadré rouge est figuré pour chacune des structures la région correspondant à la portion d'alignement réalisé par le logiciel PSI-BLAST (Figure 32).

Au plan structural (Figure 33), cette portion d'alignement correspond à la superposition de deux structures secondaires (un brin et une hélice) (Figure 33) présentes dans la région structuralement conservée entre les deux séquences. Ces observations nous

permettent d'illustrer la réalité structurale correspondant à l'homologie lointaine détectée au niveau de la séquence. Nous avons vu précédemment que les alignements du signal non significatif de PSI-BLAST présentent un taux de recouvrement très faible par rapport à la taille des séquences impliquées. Toutefois il semble que parmi ces séquences il existe des portions de petite taille présentant un taux de conservation plus important car essentielles à la structure 3D de l'ensemble ou à la fonction de la famille de protéines. Nous reviendrons à la fin du chapitre IV sur cet exemple afin d'évaluer comment fonctionne la stratégie développée dans cette thèse sur cet exemple particulier.

II.5.Conclusions

Cette étude a permis de vérifier la présence de séquences d'homologues lointains dans le signal non significatif du logiciel PSI-BLAST. De plus, nous avons pu démontrer que les alignements non significatifs calculés par le logiciel PSI-BLAST dans l'intervalle (10^{-3}, 1000) rassemblent un ensemble de séquences enrichi en homologues lointains par rapport à un échantillon aléatoire de la base de données. Nous avons par ailleurs observé que les séquences d'homologues présentes au sein du signal non significatif sont de plus en plus nombreuses lorsque la e-value augmente, mais, qu'en accord avec la définition de la e-value, leur pourcentage par rapport au nombre de séquences détectées décroît jusqu'à 1% sur l'intervalle de e-value [100-1000]. Nous avons montré que des alignements présentant des e-values fortement non significatives (>100) sont associés à une réalité structurale dans un nombre important de cas (environ un tiers). Enfin, nous avons constaté que parmi le signal non significatif du logiciel PSI-BLAST les alignements de grande taille sont généralement de mauvaise qualité ; la taille moyenne des alignements corrects à plus de 90% est de seulement trente résidus. Ces alignements corrects de faible taille peuvent correspondre à seulement quelques éléments de structure secondaire, essentiels à la structure 3D ou à la fonction des deux protéines.

Cette première analyse suggère l'importance de l'analyse du signal non significatif de PSI-BLAST et souligne l'intérêt de conserver un intervalle d'étude du signal non significatif correspondant à des e-value de 0,001 à 1000. Néanmoins, dans ces conditions, le signal non significatif est alors constitué d'un grand nombre de séquences, qui rend les expertises manuelles laborieuses à réaliser et inenvisageables à grande échelle. Ceci renforce la nécessité d'automatiser la procédure d'analyse de ce signal.

Pour la suite, j'ai donc isolé un jeu de données test de 200 séquences issues de la banque SCOP10, pour lesquelles plus de 5 séquences de la même superfamille sont retrouvées dans le signal non significatif de Psi-Blast. C'est sur ce jeu test que se fera la mise au point de ma procédure de recherche d'homologues lointains.

Chapitre III :Filtrage du signal non significatif à l'aide des méthodes de prédictions de structures secondaires.

CHAPITRE III : Filtrage du signal non significatif à l'aide des méthodes de prédiction de structures secondaires.

III.1.Introduction

L'étude réalisée au cours du chapitre II nous a montré qu'il existait des homologues lointains au sein du signal non significatif généré par le logiciel PSI-BLAST. De plus, ces homologues lointains sont six fois plus présents dans le signal non significatif que dans l'ensemble de la base de données analysée. Le signal non significatif étant plus concentré en homologues lointains que la base de données de séquences initiales, il constitue donc une base intéressante pour initier l'identification d'éventuels homologues lointains.

Toutefois, le nombre important de séquences dans le signal non significatif constitue une limite à l'exploration manuelle des séquences potentiellement intéressantes et nécessite le développement d'outils d'analyse plus performants. Nous avons donc décidé d'utiliser une première stratégie pour filtrer rapidement les séquences de ce signal en nous appuyant sur les prédictions de structure secondaire. Une analyse plus fine, par des méthodes de comparaison profil-profil plus coûteuses en temps de calcul, pourra alors être réalisée par la suite.

Les prédictions de structure secondaire constituent une source d'information structurale complémentaire dont ne tient pas compte le logiciel PSI-BLAST lors de son calcul de score d'alignement. Pourtant, comme nous l'avons décrit dans l'introduction, l'information structurale est mieux conservée au cours de l'évolution que la séquence des acides aminés. Des homologues lointains peuvent présenter des structures voisines associées à des séquences très divergentes. Dans le but de traiter l'ensemble des séquences le plus rapidement possible, les prédictions de structures secondaires sur les séquences du signal non significatif ont été effectuées sur séquences uniques sans passer par l'étape de construction préalable d'un alignement multiple . Bien que dans ce cas, la validité des prédictions soit *a priori* réduite (de 78 à 67 %) , les capacités discriminatives de cette méthode ont été explorées. Deux stratégies différentes ont été envisagées.

1 - Une stratégie de comparaison locale des structures secondaires prédites. Notre hypothèse est que, dans l'alignement restreint généré par le logiciel PSI-BLAST, les structures secondaires prédites pour la séquence d'intérêt et un éventuel homologue lointain doivent présenter une certaine similitude. Nous avons tenté d'évaluer le degré de similitude permettant de discriminer un tel homologue lointain.

2 - Une stratégie de comparaison globale des structures secondaires prédites. Dans ce cas, nous avons émis l'hypothèse que la similitude entre structures secondaires prédites doit s'étendre au-delà du segment aligné par le logiciel PSI-BLAST jusque sur l'ensemble du domaine structural associé. Le domaine structural correspondant à la séquence entière lorsque l'on travaille sur la SCOP, nous avons cherché à déterminer si la comparaison des compositions en structure secondaire des séquences potentiellement homologues permettait de déduire quelles séquences étaient de vrais homologues de la séquence d'intérêt.

III.2. Méthodes

III.2.1. Filtrage du signal non significatif à l'aide des prédictions de structures secondaires de manière locale.

Les prédictions de structures secondaires ont été réalisées à l'aide du logiciel PSI-PRED (v.2.45) . Les trois structures secondaires considérées sont l'hélice alpha (notée H), le feuillet bêta (noté E), et la structure désordonnée (notée C).

Une prédiction de structure secondaire a été réalisée sur les 200 séquences sélectionnées dans l'analyse présentée dans le chapitre II. La prédiction a été effectuée sur un alignement de la séquence construit à partir de séquences homologues identifiées dans la banque nr à l'aide du logiciel PSI-PRED (4 itérations, e-value d'inclusion égale à 5.10^{-4}). Une prédiction de structures secondaires a aussi été effectuée sur chaque séquence du signal non significatif. Cette prédiction a été réalisée sur séquence unique et non sur un alignement, car l'objectif est ici de filtrer le plus rapidement possible ce grand ensemble de séquences en évitant de passer par une recherche d'homologues. Les prédictions sont effectuées sur la région des séquences alignées dans le signal non significatif élargies de chaque coté de 30 acides aminés.

La concordance entre les prédictions de structures secondaires de chacune des séquences de référence et des séquences constituant leur signal non significatif respectif est évaluée à l'aide d'un score défini sur la Figure 34 : le Qsecpred. Ce score évalue la proportion de positions dans l'alignement pour lesquelles prédictions de structures secondaires en hélice ou brin sont identiques.

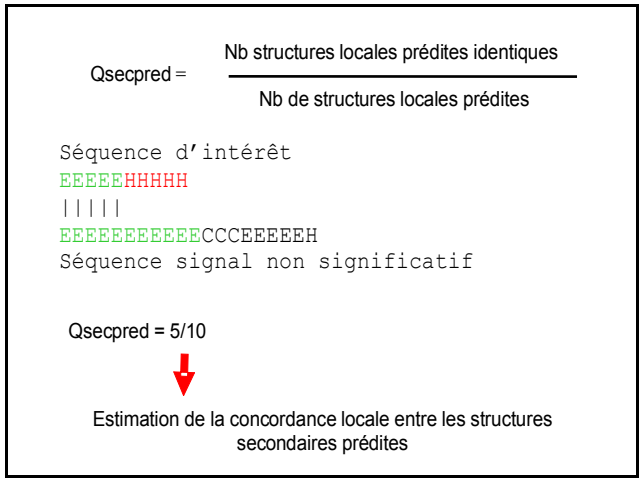


Figure 34 Schéma récapitulatif de l'approche utilisée lors du calcul du Qsecpred. Ce score est égal au rapport du nombre de positions pour lesquelles il y a accord sur l'élément de structure secondaire divisé par le nombre total de résidus de la séquence d'étude initiale. Le calcul du Qsecpred s'effectue à partir des résidus prédits comme appartenant à une hélice ou un brin.

III.2.2. Filtrage du signal non significatif à l'aide des prédictions de structures secondaires de manière globale.

Dans un premier temps, nous avons exploré quels paramètres calculés à partir des prédictions de structures secondaires permettraient de discriminer les séquences appartenant aux classes A, B, C ou D de la classification SCOP (respectivement tout hélice α, tout feuillet β, mixte α+β et mixte α/β). Pour cela nous avons étudié les capacités discriminatives de 3 facteurs, le taux d'hélices α prédites, le taux de feuillets β prédits, le taux d'alternances α/β prédits (Figure 35). Dans un premier temps, des prédictions de structures secondaires ont été effectuées avec le logiciel PSI-PRED sur un ensemble de 4387 séquences issues de la SCOP40 appartenant aux classes A, B, C et D. Ces prédictions ont été effectuées sur toute la longueur de la séquence.

Afin de comparer la composition globale en structures secondaires des séquences appartenant à chacune des classes, les pourcentages d'hélices α et de brins β ont été calculés. Pour cela, plusieurs règles de décision ont été utilisées pour définir l'existence de ces éléments de structure :

- la taille minimale pour constituer un brin est de 3 résidus consécutifs prédits en brin.
- la taille minimale pour constituer une hélice est de 4 résidus consécutifs prédits en hélice.

– la taille minimale pour considérer l'apparition d'une boucle et ainsi séparer deux éléments de structure secondaire est de 4 résidus consécutifs prédits comme appartenant à une structure désordonnée.

Afin de discriminer entre les repliements de type α+β et α/β, le pourcentage d'enchaînements de deux éléments de structure secondaire identiques, appelé indice de succession, a été calculé sur l'ensemble de la séquence (Figure 35). Pour les structures de type α+β on suppose que l'indice de succession tend vers 100 % (nombreux enchaînements de structures secondaires de même nature) tandis que pour des structures avec une alternance fréquente de structures secondaires telles que les α/β, cet indice tend vers 0 % (chaque hélice est suivie d'un brin et inversement).

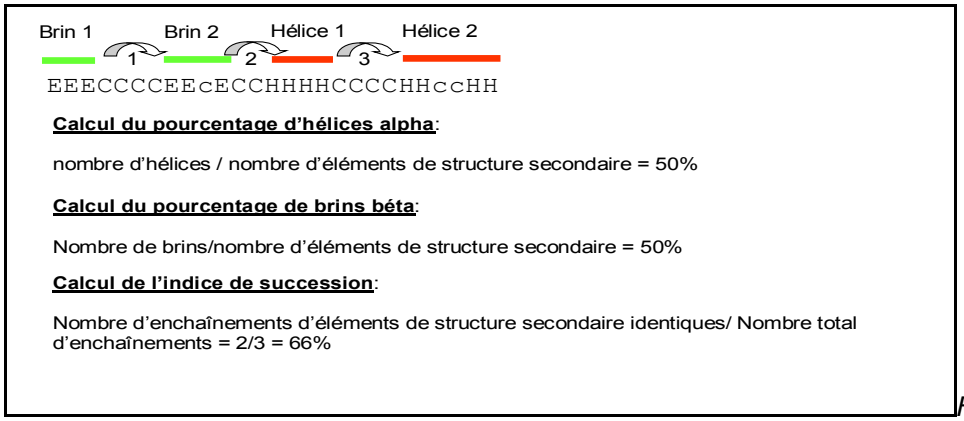


Figure 35 : Schéma descriptif des principes de calcul des différents paramètres utilisés pour la comparaison des prédictions de structures secondaires de manière globale.

Les résultats de cette analyse globale ont permis de définir des seuils pour chacun des trois paramètres présentés Figure 35. Ces seuils ont ensuite été utilisés pour évaluer si la prédiction de structures secondaires effectuée sur la globalité des séquences identifiées dans le signal non significatif pouvait permettre de filtrer les homologues lointains.

Pour cela, une prédiction de structures secondaires a été réalisée sur l'alignement multiple associé à chacune des 200 séquences de notre étude calculé avec logiciel PSI-PRED (4 itérations sur la banque nr, e-value d'inclusion de 10^{-4}). Ensuite, pour chacune des séquences du signal non significatif, les séquences ont été récupérées dans leur intégralité et une prédiction de structures secondaires sur séquence unique a été effectuée avec le logiciel PSI-PRED. Les taux d'hélices α, de feuillets β et les indices de successions ont été calculés sur la globalité des séquences et comparés pour effectuer l'étape de filtrage.

III.3.Résultats

III.3.1.Utilisation des structures secondaires d'un point de vue local

Afin d'étudier le potentiel discriminant des prédictions de structures secondaires d'un point vue local, les valeurs de Qsecpred ont été calculées pour l'intégralité des alignements présents dans le signal non significatif généré par le logiciel PSI-BLAST pour nos 200 cas d'étude.

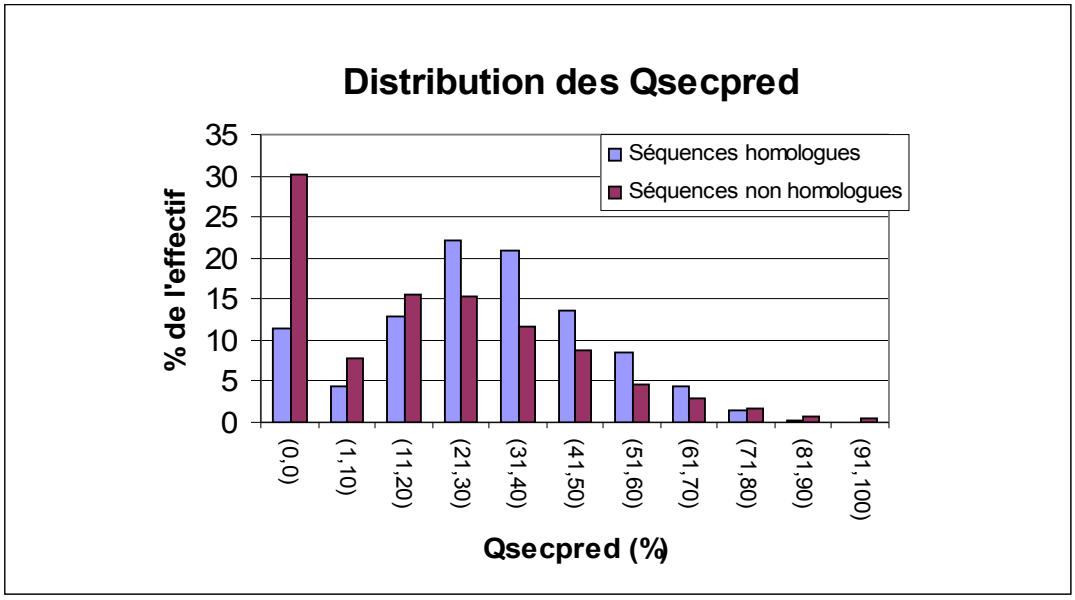


Figure 36 : Distribution des Qsecpred des séquences homologues par rapport aux séquences non homologues du signal non significatif. En ordonné figure le nombre d'alignement, en abscisse les intervalles de Qsecpred (de 0 à 100% par intervalle de 10). La distribution des séquences non homologues est représentée par l'histogramme rouge, la distribution des séquences homologues par l'histogramme bleu.

Sur la Figure 36, la distribution des Qsecpred obtenus avec les séquences d'homologues lointains est comparée à la distribution des Qsecpred obtenus avec les autres séquences. Sur ce graphique, la distribution des Qsecpred associée aux séquences non homologues du signal non significatif présente un maximum pour les Qsecpred nuls. La distribution présente un second maximum pour la valeur de Qsecpred de 20 %. La comparaison des effectifs cumulés nous permet d'observer que 50% des séquences non homologues présentent des Qsecpred inférieurs à 20%. Il est intéressant d'observer que pour des Qsecpred supérieurs à 20 %, les tendances des deux distributions sont relativement semblables. Ces résultats nous informent qu'une proportion importante des alignements entre séquences étudiées et séquences non homologues du signal non significatif est localisée au

niveau d'éléments de structures secondaires prédits comme identiques à ceux de la séquence étudiée.

La distribution des Qsecpred associée aux séquences homologues du signal non significatif s'organise de manière croissante puis décroissante autour d'une valeur maximale de Qsecpred de 30%. Nous observons aussi que le taux de séquences présentant un Qsecpred nul est très inférieur à celui observé pour des séquences non homologues (environ 20% de l'effectif total en moins). De plus, une étude approfondie des séquences homologues présentant un Qsecpred compris entre 0 % et 10 % montre que 90% des séquences sont associés à un Qmod nul (Figure 25 du chapitre II), c'est-à-dire que 90% des séquences présentant un Qsecpred de 0 % à 10 % sont associées à un alignement fortuit ne présentant aucune réalité structurale. De manière générale, ce type d'alignement ne présente que peu d'intérêt pour la suite de notre étude et ne constitue pas en soit une perte d'information importante.

Ainsi, bien que le décalage entre les deux distributions soit faible, la proportion d'alignements fortuits observés dans l'intervalle de 0 à 20 %, permet de proposer une valeur seuil de filtrage entre séquences homologues et non homologues dans le signal non significatif. La valeur seuil sélectionnée correspond à un Qsecpred de 20 %. Pour évaluer la capacité de filtrage du signal non significatif de PSI-BLAST avec ce seuil de Qsecpred, nous avons calculé, pour nos 200 cas d'étude, le nombre moyen de séquences homologues et non homologues dans le signal non significatif avant et après filtrage.

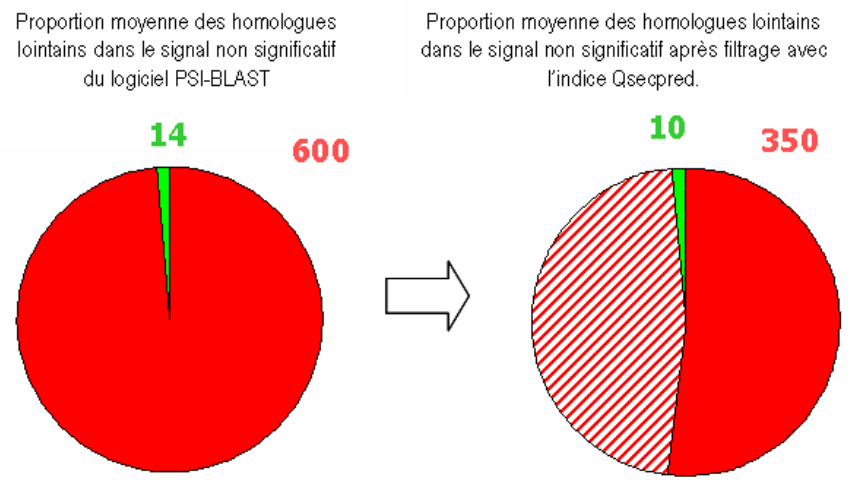


Figure 37 : Diagramme représentant la proportion de séquences homologues et non homologues filtrées pour des valeurs de Qsecpred supérieur à 20 %. En vert est figurée la proportion de séquences d'homologues lointains présentes dans le signal non significatif, en rouge la proportion de séquences non homologues, en hachuré la proportion de séquences non homologues filtrées.

Sur les graphiques (Figure 37) on observe qu'initialement les séquences homologues représentent environ 1% des séquences constituant le signal non significatif. Après filtrage, le nombre de séquences non homologues est réduit d'environ 40% alors qu'une perte de séquences homologues de l'ordre de 25% est observée. Rappelons que parmi ces 25 %, la plupart des séquences possèdent un Qmod nul avec la séquence de référence.

Ainsi, cette approche permet d'éliminer un nombre conséquent de séquences non homologues du signal non significatif. L'élimination reste néanmoins insuffisante pour identifier directement les homologues lointains. En revanche, cette réduction du nombre de séquence est susceptible de réduire fortement le nombre de constructions de profils requis pour une étude approfondie basée sur les méthodes de comparaison profil/profil.

III.3.2. Utilisation des structures secondaires sur les séquences globales.

Notre hypothèse de travail est qu'en appliquant les prédictions de structures secondaires à la globalité des séquences, il est possible d'identifier à quelle classe de la classification SCOP une séquence appartient. Ainsi, on peut imaginer que si une séquence d'intérêt peut être associée à l'une des quatre classes A, B, C et D (respectivement tout hélice α, tout feuillet β, mixte α+β et mixte α/β), il est possible d'éliminer du signal non significatif toutes les séquences prédites comme appartenant à une classe différente.

Nous avons analysé chaque classe indépendamment et évalué quel paramètre était le mieux adapté pour éliminer du signal non significatif le maximum de séquences n'appartenant pas à cette classe. Dans un premier temps, les classes A et B ont été étudiées en utilisant les taux d'hélice α et de feuillet β pour caractériser leur appartenance à l'une ou l'autre de ces classes.

I.1.1.14Utilisation du taux d'hélice α

Pour chacune des 4387 séquences issues de la banque SCOP40, le taux d'hélice α a été calculé à partir de la prédiction de structures secondaires réalisée sur la séquence entière du domaine. Les distributions des taux d'hélice α dans les séquences appartenant à la classe A ont été ensuite comparées au taux d'hélices retrouvé chez les séquences n'appartenant pas à la classe A.

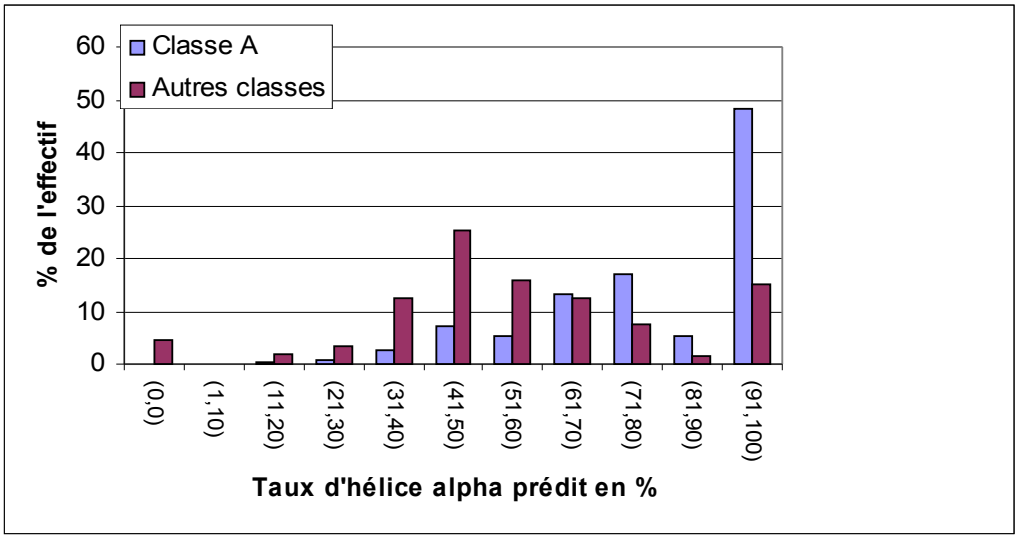


Figure 38 Distribution du taux d'hélice alpha prédit en fonction de l'appartenance ou non à la classe A. En abscisse figure le taux d'hélice alpha (de 0 a 100 % par intervalle de 10 %), en ordonné les pourcentages de séquences au sein de l'effectif appartenant à chaque intervalle. En bleu figure la distribution des séquences associées à la classe A. En rouge figure la distribution des séquences n'étant pas associées à la classe A. L'étude est réalisée sur 911 séquences appartenant à la classe A et 3476 séquences regroupant les séquences des classes B, C et D.

Le graphique Figure 38 montre qu'il existe une différence au niveau des distributions du taux d'hélice α des séquences de la classe A et de l'ensemble des autres classes. Les séquences appartenant à la classe A suivent une distribution croissante avec un minimum prédit d'environ 30% d'hélice α et un maximum (50% des cas) pour des taux d'hélice α compris entre 90 et 100 %. En revanche, les séquences n'appartenant pas à la classe A suivent une distribution centrée autour d'un maximum (25% des cas) pour des taux d'hélice α compris entre 40 et 50 %.

Ces données montrent que lorsqu'une séquence présente un taux d'hélice α prédit supérieur à 70 % il est possible de lui assigner un repliement de type pur hélice α (classe A). Un tel filtre permet d'éliminer 75 % des séquences n'appartenant pas à la classe A tout en conservant 70 % appartenant à la classe. Il est envisageable d'utiliser ce seuil afin de filtrer le signal non significatif des séquences présentant un taux d'hélice supérieur à 70%.

I.1.1.15Utilisation du taux de feuillets β.

En suivant la même approche que celle décrite précédemment, la distribution du taux de feuillets β a été calculée et comparée au taux de feuillet β des séquences n'appartenant pas à la classe B.

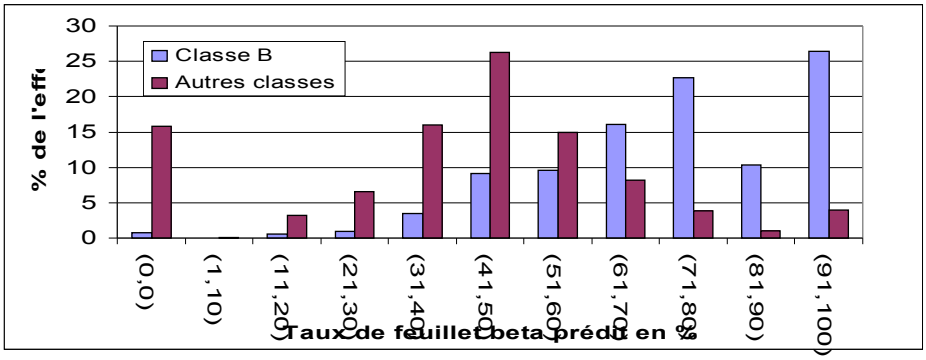


Figure 39 : Distribution du taux de feuillet β prédit en fonction de l'appartenance ou non à la classe B. En abscisse figure le taux de feuillet β (de 0 a 100 % par intervalle de 10 %), en ordonnée les pourcentages de séquence au sein de l'effectif appartenant à chaque intervalle. En bleu figure la distribution des séquences associées à la classe B. En rouge figure la distribution des séquences n'étant pas associées à la classe B. L'étude est réalisée sur 1093 séquences issues de la classe B et 3294 séquences regroupant les séquences des classes A, C et D.

Le graphique de la Figure 39 montre que la distribution des taux de feuillets β prédits dans les séquences de la classe B est décalée par rapport à celle observée pour les autres classes. Comme précédemment le choix d'un taux de feuillets β seuil de 70 % pour conclure à l'appartenance à la classe B semble bien adapté. Au-dessus de ce seuil, on peut estimer que la séquence d'intérêt et celle du signal non significatif adoptent un repliement en pur feuillet β. Ce seuil permet de récupérer environ 60% des séquences appartenant à la classe B tout en éliminant environ 90% des séquences n'appartenant pas à la classe B.

I.1.1.16Analyse de l'indice de succession des structures secondaires dans les classes C et D.

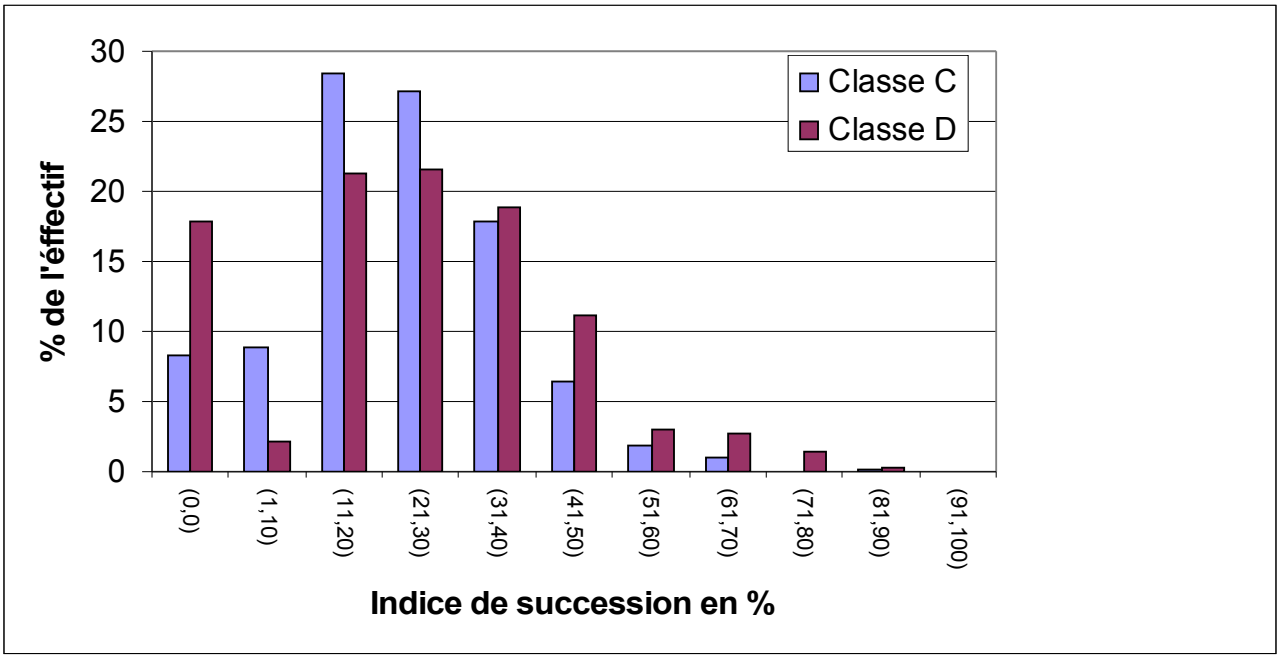


Figure 40 : Distribution de l'indice de succession. En abscisse figure le pourcentage d'enchaînements de deux éléments de structure secondaire identiques (de 0 a 100 % par intervalle de 10 %), en ordonnée le pourcentage de séquences associées aux intervalles d'indices de succession. En rouge figure la distribution des séquences associée à la classes C (α+β) , en bleu figure la distribution des séquences appartenant à la classe D (α/β). L'étude est réalisée sur 11O1 séquences de la classe C et 1282 séquences de la classe D.

L'indice de succession a été calculé selon la méthode décrite sur la Figure 35 pour les séquences appartenant aux classes C et D parmi les 200 séquences de notre échantillon d'étude et les séquences appartenant au signal significatif et non significatif du logiciel PSI-BLAST. Nous avons vu que les taux d'hélices α et de feuillets β pouvaient être en mesure de séparer en partie les classes A et B des autres classes présentes parmi un ensemble de séquences. Toutefois ces résultats ne s'appliquent pas aux compositions mixtes α+β et α/β des structures secondaires associées aux séquences appartenant aux familles C et D. Nous avons donc envisagé d'utiliser l'indice de succession (Figure 35) afin de séparer ces deux classes.

Sur le graphique Figure 40 nous pouvons voir qu'il est difficile de discerner un décalage entres les séquences appartenant à la classe C (α+β) et D (α/β). En effet, les distributions de l'indice de succession dans les classes C et D suivent des distributions centrées autour des indices de successions compris entre 20 et 40% : environ 27% des cas pour les séquences appartenant à la classe C pour 21% des cas appartenant à la classe D. La

superposition globale des deux distributions ne permet pas d'envisager une séparation des séquences des classes C et D sur un critère tel que l'indice de succession.

I.1.1.17Evaluation de la récupération des homologues du signal non significatif à l'aide de l'utilisation des prédictions de structures secondaires d'un point de vue global.

Pour la suite de cette étude, nous avons évalué les seuils permettant de distinguer les classes A, B, C et D. Les seuils utilisés sont les suivants et sont répertoriés dans la Figure 41 :

- une séquence appartient à la classe A, si elle présente un taux d'hélices α prédit supérieur à 70%
- une séquence appartient à la classe B, si elle présente un taux de feuillets β prédit supérieur à 70%
- une séquence appartient à la classe C ou D si elle présente des taux de feuillets β et d'hélices α inférieurs à 70%.

L'indice de succession n'a pas été repris au vu de ses faibles performances discriminatives.

Brin β > 70%	Brins β < 70% et helices α < 70%	Hélice α > 70%
Classe B	Classe C ou D	Classe A

Figure 41 : Tableau récapitulatif des seuils de pourcentage de structure secondaire employés afin de filtrer les séquences du signal non significatif.

Ces valeurs seuils sont utilisées afin de réaliser un filtrage du signal non significatif. Ce filtrage aura pour objectif de récupérer uniquement les séquences appartenant à la classe de la séquence étudiée. La qualité du filtrage est évaluée pour les séquences appartenant aux classes A, B, C et D parmi les 200 séquences de notre étude.

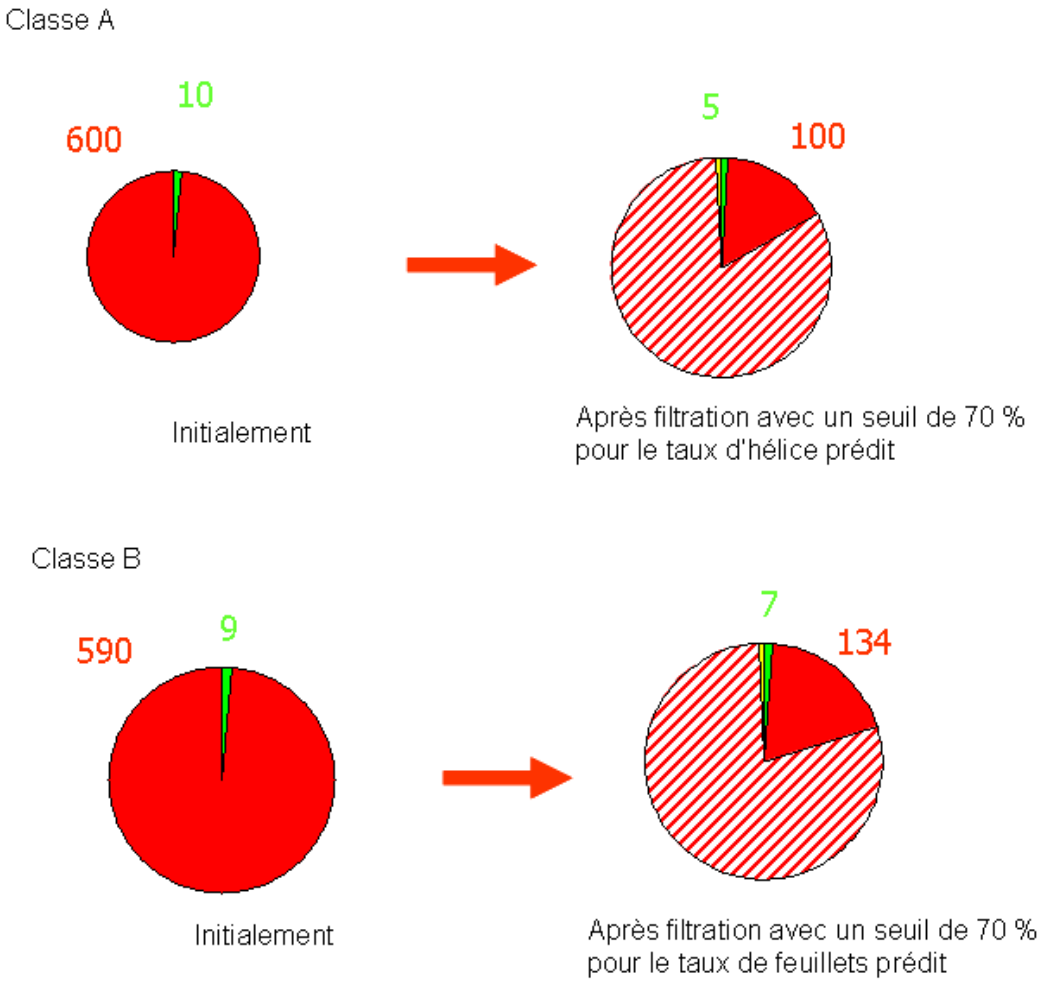


Figure 42 : Diagramme représentant la proportion de séquences homologues et non homologues filtrées pour les séquences appartenant à la classe A (en haut) et la classe B (en bas). En vert, le nombre moyen de séquences homologues présentes dans le signal non significatif, en rouge, le nombre moyen de séquences non homologues dans ce signal, en hachuré le nombre moyen de séquences non homologues éliminées. Les chiffres verts représentent le nombre moyen de séquences homologue, les chiffres rouges le nombre moyen de séquences homologues. Enfin, en jaune, le nombre moyen de séquences homologues éliminées à tord. Cette étude porte sur 29 séquences appartenant à la classe A et 28 séquences de la classe B au sein des 200 cas d'études.

Les résultats du filtrage par le taux d'hélices α ou de feuillets β montrent que pour les séquences prédites avec un taux de structures secondaires supérieur à 70%, le filtrage du signal non significatif permet d'éliminer environ 80% du signal non significatif. Néanmoins, cette sélection entraîne la perte d'environ 50% des séquences d'homologues lointains appartenant au signal non significatif de la classe A et 33% des séquences de la classe B (Figure 42). On peut suggérer que cette perte de séquences homologues après filtrage est due à l'existence d'une variabilité structurale trop importante entre certains homologues lointains (insertion de boucles contenant des éléments de structure secondaire) pour permettre leur

détection en s'appuyant sur l'analyse des prédictions de structures secondaires calculées de façon globale.

La Figure 43 présente les résultats du filtrage par l'utilisation couplée du taux d'hélice α et de brins β pour sélectionner les séquences appartenant à la classe C ou D. On observe que pour les séquences présentant initialement des taux de feuillets et d'hélices inférieurs à 70%, la filtration du signal non significatif ne permet d'éliminer que 30% du signal non significatif ; la méthode présente alors une spécificité moins importante que celle obtenue lors de l'utilisation des prédictions de secondaires de manière locale. La perte des séquences homologues présentes au sein du signal non significatif est de l'ordre de 20%. Ainsi, l'utilisation des prédictions des taux d'hélice α et de feuillets β afin de filtrer le signal non significatif ne permet pas d'obtenir une spécificité intéressante pour les séquences appartenant aux classes C et D.

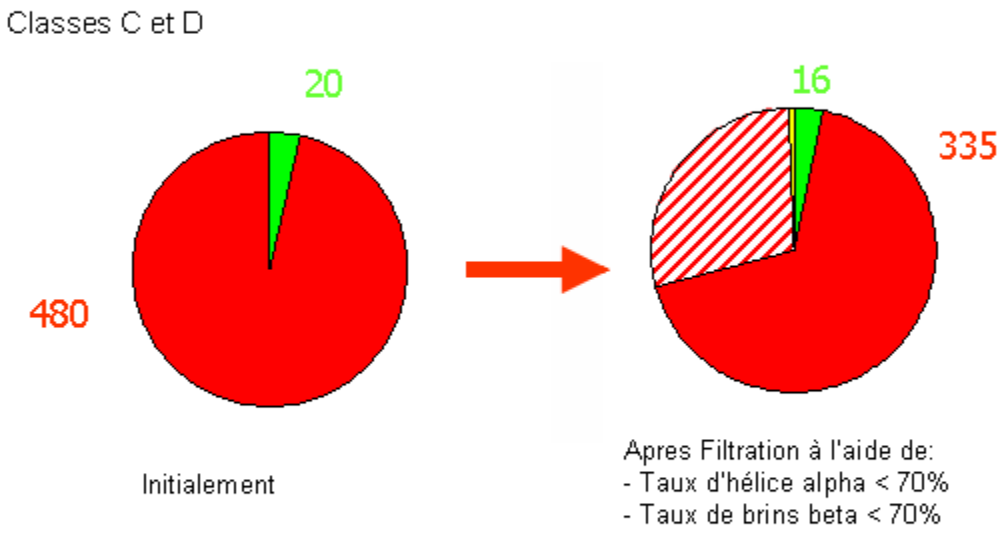


Figure 43 Diagramme représentant la proportion de séquences homologues et non homologues filtrées pour les séquences appartenant aux classes C et D. En vert est figurée la proportion de séquences homologues présentes dans le signal non significatif, en rouge la proportion de séquences non homologues, en hachuré la proportion de séquences non homologues filtrées. En jaune le pourcentage de séquences homologues filtrées. Les chiffres verts représentent le nombre moyen de séquences homologues, les chiffres rouges le nombre moyen de séquences non homologues. Cette étude porte sur 80 séquences appartenant aux classes C et D au sein des 200 cas d'études.

III.3.3. Etude de la complémentarité des stratégies locales et globales pour le filtrage du signal non significatif réalisé avec les prédictions de structures secondaires.

Pour conclure cette analyse un croisement des résultats obtenus dans les sections précédentes a été effectué pour explorer la complémentarité des stratégies de prédiction de structures secondaires locales et globales. Cette étude a été réalisée pour les séquences appartenant aux classes A et B et présentant respectivement un taux d'hélice α et de feuillet β supérieur à 70 %. Les séquences appartenant aux classes C et D et présentant un taux d'hélice α et de feuillet β inférieur à 70 % n'ont pas été retenues pour cette analyse complémentaire à cause du peu de spécificité apportée par la filtration par l'approche globale.

Pour les séquences appartenant à la classe A

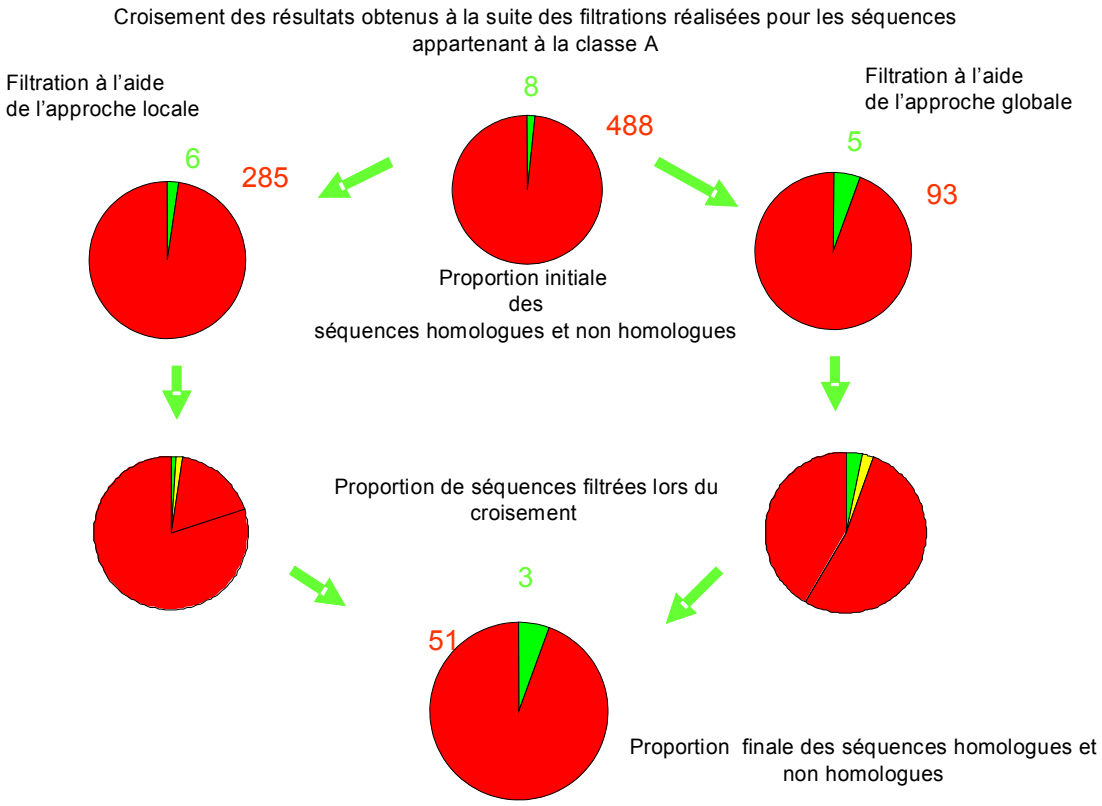


Figure 44 Diagramme représentant les proportions de séquences homologues et non homologues filtrées pour les séquences appartenant à la classe A. Au sein des diagrammes : en vert est figurée la proportion de séquences homologues présentes dans le signal non significatif, en rouge la proportion de séquences non homologues, en hachuré la proportion de séquences non homologues filtrées, en jaune le pourcentage de séquences homologues filtrées. Les chiffres verts représentent le nombre moyen de séquences homologue, les chiffres rouges le nombre moyen de séquences homologues. Pour cette étude 29 séquences appartenant à la classe A sont analysées.

Pour les séquences appartenant à la classe B

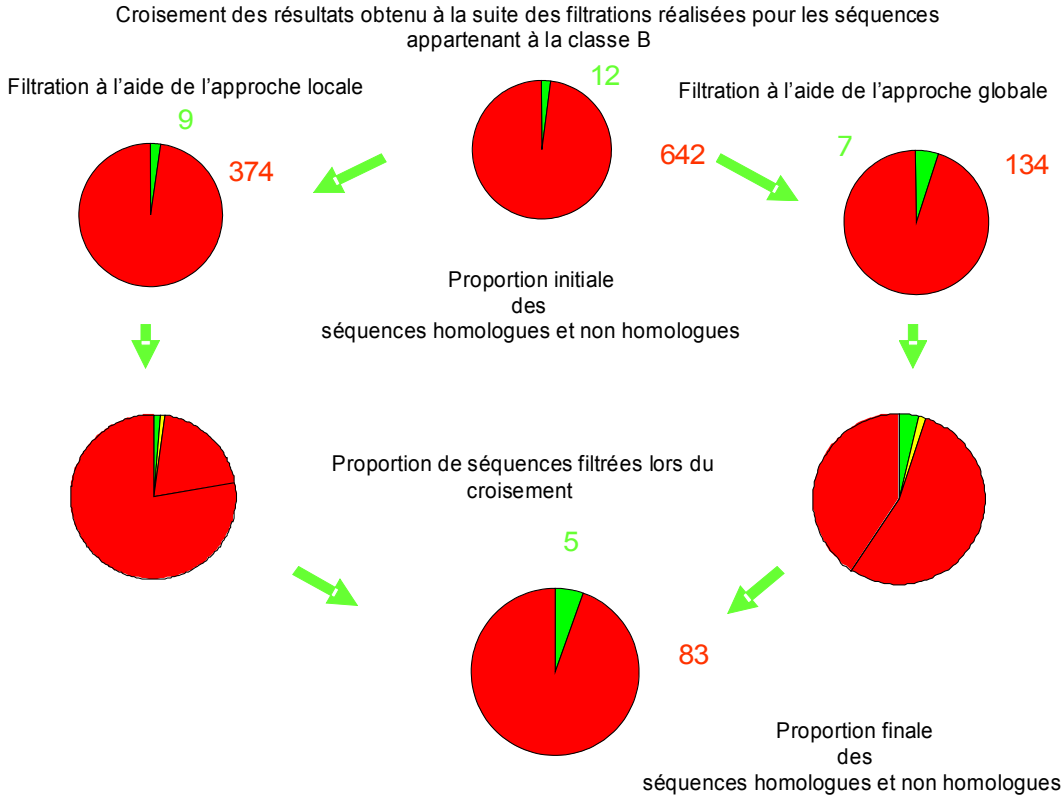


Figure 45 : Diagramme représentant la proportion de séquences homologues et non homologues filtrées pour les séquences appartenant à la classe B. En vert est figurée la proportion de séquences homologues présentes dans le signal non significatif, en rouge la proportion de séquences non homologues, en hachuré la proportion de séquences non homologues filtrées. En jaune le pourcentage de séquences homologues filtrées. Les chiffres verts représentent le nombre moyen de séquences homologues, les chiffres rouges le nombre moyen de séquences non homologues. Pour cette étude 28 séquences appartenant à la classe B sont analysées.

Le croisement des données des approches locales et globales pour les séquences appartenant à la classe A (Figure 44) et à la classe B (Figure 45) permet d'éliminer 80% des séquences non homologues présentes lors de la filtration du signal non significatif à l'aide d'une approche locale. De plus cette approche permet également d'optimiser le filtrage à l'aide de l'approche globale en éliminant environ 45% des séquences non homologues présentes. Cependant ce double criblage entraîne au final une perte moyenne de 55% des homologues présents au sein du signal. Pour les séquences de la classe A, ce double filtrage entraîne ainsi en moyenne la perte de 5 séquences homologues (~60% de perte). Pour les séquences de la classe B : 7 séquences homologues (~60% de perte)

Le croisement de ces données entraîne une perte non négligeable des homologues présents dans le signal non significatif. Cette perte est trop importante au regard de la faible

élimination des séquences non homologues. L'intérêt d'un tel croisement apparaît donc discutable dans le cadre d'une recherche d'homologues lointains.

III.4. Conclusions

Au cours de ce chapitre nous avons vu que les prédictions de structures secondaires utilisées avec une stratégie locale ou globale se révèlent tout à fait intéressantes pour filtrer le signal non significatif. L'approche globale présente des résultats intéressants pour les domaines présentant une organisation en structures secondaires prédites à plus de 70% en hélices α ou en feuillets β. En dehors de ces conditions, cette stratégie n'apporte pas de performances intéressantes. De plus, les tests réalisés pour l'étude de l'approche globale sont effectués sur une banque de séquences de domaines; l'utilisation de cette approche sur une banque de séquences comportant plusieurs domaines pose des difficultés pour définir la zone sur laquelle doit être effectuée la prédiction de structure secondaire en vue d'une analyse « globale ». Pour la suite de cette étude nous avons donc préféré l'utilisation d'une approche locale des prédictions de structures secondaires. Cette approche, parfois moins efficace que l'approche globale en termes de filtrage, permet toutefois d'être utilisée sur n'importe quel type de séquences. Elle permet un filtrage correct d'environ 40 % des séquences non homologues du signal non significatif. Par ailleurs, nous avons observé que la majorité des séquences homologues perdues lors de la filtration était associée à un alignement structural fortuit. Enfin, cette étape est rapide à calculer : une prédiction de structure secondaire sur séquence unique réalisée sur le signal non significatif d'une sortie PSI-BLAST composée de 600 séquences ne prend que quelques minutes.

L'une des perspectives de ces travaux est d'utiliser des méthodes de prédictions de structure secondaire plus fiables. En effet, les prédictions de structure secondaire effectuées sur séquence unique par le logiciel PSI-PRED présentent une fiabilité limitée (68 %), et des méthodes plus récentes (cf chapitre VI) devraient permettre de mieux discriminer les vrais homologues structuraux des faux positifs de PSI-BLAST.

De manière générale, le filtrage par les deux méthodes de prédiction de structure secondaire proposées n'est pas assez efficace pour réaliser aisément une analyse manuelle approfondie de l'ensemble des séquences selectionnées en vue de détecter des homologues lointains. Afin de gagner en capacité de filtrage, nous avons exploré les potentialités des méthodes de comparaison profil/profil.

Chapitre IV :Filtrage du signal non significatif à l'aide des méthodes de comparaison profil/profil

IV.1. Introduction

L'identification d'homologues lointains au sein des séquences alignées de façon non significative par le logiciel PSI-BLAST nécessite d'intégrer des informations supplémentaires pour éliminer les alignements fortuits et sélectionner ceux potentiellement intéressants. L'étude réalisée dans le chapitre III a montré que l'utilisation des prédictions de structure secondaire permettait de réaliser un filtrage du signal non significatif réduisant de manière conséquente le nombre de séquences non homologues tout en conservant pratiquement l'intégralité des homologues lointains. Néanmoins, cette approche rapide reste peu spécifique et nécessite d'être couplée à une détection plus stricte pour présenter un intérêt pratique.

Dans ce chapitre, nous présentons les résultats obtenus avec une seconde stratégie s'appuyant sur les comparaisons profil/profil à l'aide des logiciels COMPASS et HHsearch. Comme décrit dans l'introduction, ces deux méthodes sont basées sur des algorithmes très différents. Elles présentent des performances tout à fait intéressantes individuellement susceptibles d'être encore améliorées par une exploitation conjointe des résultats. Pour chaque séquence alignée dans le signal non significatif de PSI-BLAST avec une des 200 séquences de la base test générée au chapitre II, un profil a été généré (Figure 46). Chaque profil a ensuite été comparé au profil associé à la séquence de référence et nous avons ensuite estimé les valeurs de scores pour lesquelles la discrimination entre homologues lointains et non homologues était effective. La construction des profils constitue une étape coûteuse en temps de calcul et nous avons donc exploré dans la suite comment un couplage entre les approches rapides de prédiction de structure secondaire et les comparaisons profil/profil pourrait améliorer l'efficacité de la recherche.

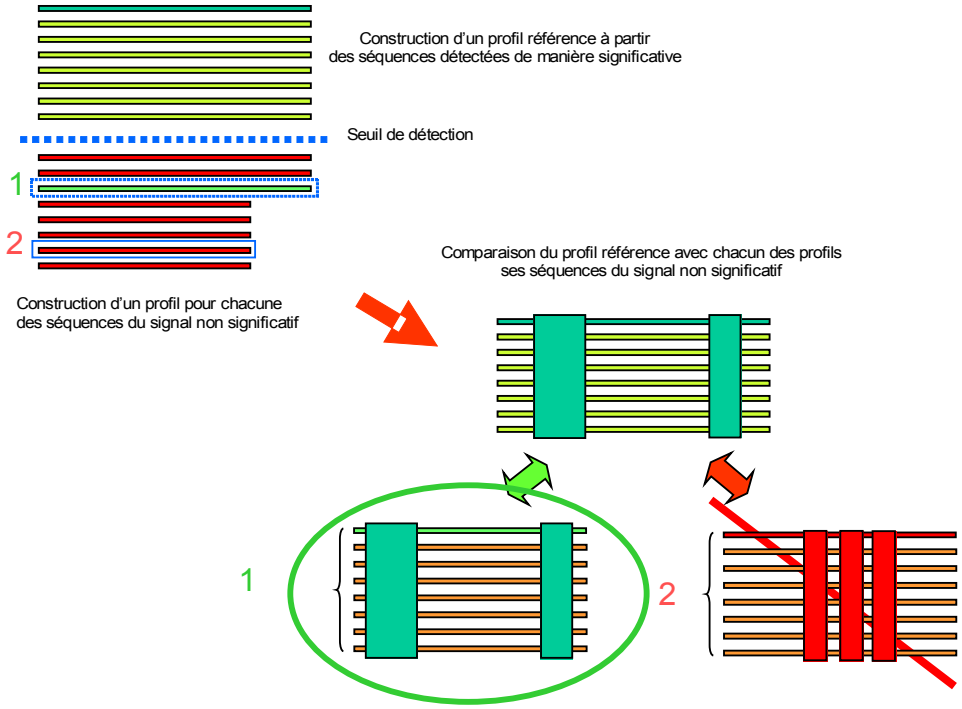


Figure 46 : Schéma récapitulatif de la filtration du signal non significatif à l'aide des méthodes de comparaison profil à profil.

IV.2.Méthodes

IV.2.1.Construction des alignements de références.

Pour chacune des 200 séquences de référence (cf chapitre II), une recherche de séquences homologues a été effectuée avec le logiciel PSI-BLAST (v. 2.2.13) (4 itérations, e-value d'inclusion de 5.10^{-4}) sur la banque de données de séquence nr (septembre 2004). Les alignements avec chacune des séquences détectées présentant une e-value inférieure à 10^{-3} sont inclus dans l'alignement de référence. Les alignements par paires sélectionnés sont ensuite convertis de manière à former un alignement multiple global de référence sans l'emploi d'une étape d'alignement multiple additionnelle (utilisationde l'option « query anchored » du logiciel PSI-BLAST). Les 200 alignements ainsi générés sont désignés dans la suite par le terme d'alignements de référence.

IV.2.2. Construction des alignements tests.

Comme nous l'avons vu au cours du chapitre II, les alignements proposés parmi les séquences du signal non significatif sont généralement de petite taille. Afin que la recherche d'homologues ne soit pas limitée par la taille des fragments alignés, 30 résidus ont été ajoutés aux extrémités des fragments alignés dans le signal non significatif. Pour chacun des fragments de séquences «élargis» présents dans l'intervalle de e-value non significatif du logiciel PSI-BLAST [10^{-3};1000] de chacune des séquences référence, une recherche de séquences homologues a été effectuée sur la banque nr (septembre 2004) avec le logiciel PSI-BLAST (v. 2.2.13) (4 itérations, e-value d'inclusion égale à 5.10^{-4}). Les alignements par paires pour les séquences présentant des e-value inférieures à 10^{-3} sont convertis en alignement multiple. L'ensemble des alignements réalisés pour chacune des séquences du signal non significatif sera défini sous le terme d'alignement test.

IV.2.3.Prédiction des structures secondaires.

Pour chaque séquence référence et chaque séquence test du signal non significatif une prédiction de structure secondaire a été effectuée à l'aide du logiciel PSI-PRED (v. 2.45). Ces prédictions ont été effectuées à partir d'un alignement de séquences multiple généré par logiciel PSI-PRED (4 itérations, e-value d'inclusion égal à 10^{-4}). Ces prédictions ont ensuite

été intégrées dans les fichiers d'entrée du logiciel HHsearch afin de tester l'apport de la prédiction de structure secondaire.

IV.2.4. Construction des profils et comparaison profil/profil.

Afin de réaliser une analyse comparative nous avons employé les deux logiciels de comparaison profil/profil : COMPASS (v. 1.24) et HHSearch (v. 1.2.0) avec les options standard des deux programmes. Lors de l'utilisation du logiciel COMPASS, il n'est pas nécessaire d'effectuer un post-traitement suite à la construction des alignements. Les deux alignements au format CLUSTAL sont directement exploitables par le logiciel.

L'utilisation du logiciel HHsearch, nécessite plusieurs étapes de traitement suite à la construction des alignements. Pour obtenir des e-values correctes, il est nécessaire de calibrer le profil HMM requête sur la base de données regroupant un échantillon représentatif de profils issus de la banque SCOP. Pour les prédictions de structures secondaires, elles ont été intégrées dans l'en-tête du fichier HMM en respectant la position des insertions dans la première séquence du profil. Les trois états prédits (H, E, C) ont été intégrés ainsi que les indices de confiance calculés par PSI-PRED. Dans le cas où les structures secondaires sont prises en compte, il est nécessaire de calibrer à nouveau le profil HMM. Pour chaque séquence référence, l'ensemble des profils tests sont regroupés dans un même fichier pour former une base de données contenant l'ensemble des profils des séquences détectées de façon non significative pour cette séquence.

Pour le logiciel COMPASS, la comparaison s'effectue par paire entre le profil référence et chacun de ces profils tests respectifs. Pour le logiciel HHsearch, le profil référence est comparé à l'ensemble de la base de profils tests.

IV.2.5.Filtrage du signal non significatif.

Deux stratégies ont été testées pour filtrer le signal non significatif.

(i) Utilisation des comparaisons profil/profil uniquement.

Pour chacune des séquences présentes parmi les résultats non significatifs des requêtes PSI-BLAST, une comparaison profil/profil est effectuée entre le profil de la séquence d'intérêt et le profil d'une séquence du signal non significatif (Figure 46). Les scores calculés par les logiciels COMPASS et HHsearch doivent permettre de fixer un seuil de décision

séparant les détections fortuites du logiciel PSI-BLAST des séquences présentant potentiellement une relation d'homologie lointaine.

(ii) Utilisation des méthodes de prédiction de structures secondaires suivies des comparaisons profil/profil.

Cette stratégie vise à accélérer le processus d'analyse du signal non significatif. L'objectif est d'évaluer si l'utilisation des prédictions de structure secondaire en amont des comparaisons profil/profil est capable de pré-filtrer efficacement le nombre de profils à comparer. Pour cela, nous utilisons le paramètre Qsecpred, introduit au chapitre III et calculé à partir de l'alignement local des prédictions de structure secondaires. Les critères utilisés pour effectuer cette filtration ont été définis dans le chapitre III :

- Les séquences présentant un Qsecpred inférieur à 20 % sont considérées comme non homologues à la séquence étudiée.

- Les séquences présentant un Qsecpred supérieur à 20 % sont considérées comme des homologues lointains potentiels du signal non significatif et seront étudiées par la suite à l'aide d'une approche profil/profil.

IV.3. Résultats

IV.3.1. Qualité des alignements et capacité discriminante des logiciels COMPASS et HHsearch pour le traitement du signal non significatif de PSI-BLAST.

I.1.1.18. Evaluation de la qualité des alignements des séquences homologues calculés par le logiciel COMPASS.

Après qu'un profil ait été construit pour chaque séquence du signal non significatif, il est comparé au profil calculé à partir de la séquence d'intérêt à l'aide du logiciel COMPASS. Dans la suite de cette analyse, nous avons évalué d'une part, la qualité des alignements fournis par COMPASS, et d'autre part, la capacité discriminante des scores associés à chacun de ces alignements dans le but de repérer les séquences homologues du signal non significatif. Nous avons tout d'abord évalué la qualité locale des alignements en comparant les valeurs du paramètre Qmod (définis au chapitre II) obtenu avec les logiciels PSI-BLAST et COMPASS (Figure 47).

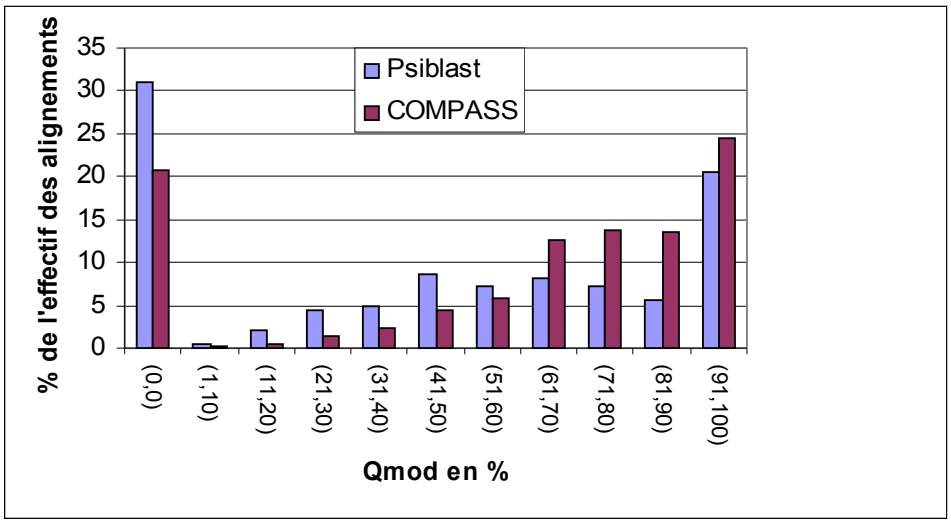


Figure 47 : Distributions des Qmod pour les alignements réalisés par le logiciel COMPASS et Psiblast à partir des séquences du signal non significatif du logiciel PSI-Blast. En abscisse figurent les intervalles de Qmod. En ordonnée figure le nombre d'alignements associés à un Qmod donné. En bleu figurent les résultats du logiciel PSI-Blast. En rouge figurent les résultats du logiciel COMPASS.

Sur la Figure 47, les valeurs de Qmod obtenus avec le logiciel COMPASS (histogramme rouge), atteignent deux maxima : un pic pour les alignements présentant une valeur de Qmod nulle (20% des cas) ainsi qu'un 2^{e} pic pour des valeurs de Qmod comprises entre 90 et 100 (25% des cas). Si on cumule l'ensemble des cas présentant un Qmod non nul, on voit que le nombre d'alignements présentant une valeur de Qmod non nul représente ~80%

des cas. Passé le premier pic correspondant à un Qmod nul, la suite de la distribution des Qmod est croissante. 90 % des alignements possédant un Qmod non nul atteignent une précision assez élevée, correcte sur plus de 50 % des positions alignées.

En comparaison des Qmod associé aux alignements calculés par le logiciel PSI-BLAST (histogramme bleu), les résultats de COMPASS présentent une nette amélioration :

- le pourcentage d'alignements possédant un Qmod nul est de 10% inférieur dans le cas de COMPASS. Ainsi, pour ~33% des alignements PSI-BLAST pour lesquels l'alignement était entièrement décalé (Qmod nul), le logiciel COMPASS parvient à réajuster partiellement l'alignement local de manière exacte. Ce réalignement peut être attribué à plusieurs facteurs, (i) un réalignement au sein même du fragment de séquence qui était aligné dans le signal significatif, (ii) un réalignement correct en dehors du fragment aligné, dans les extensions ajoutées pour élargir la longueur de la séquence test (30 résidus de chaque coté).

- il existe une différence de distribution entre les deux distributions (bleue et rouge) suggérant que la qualité des alignements augmente sensiblement avec le programme COMPASS.

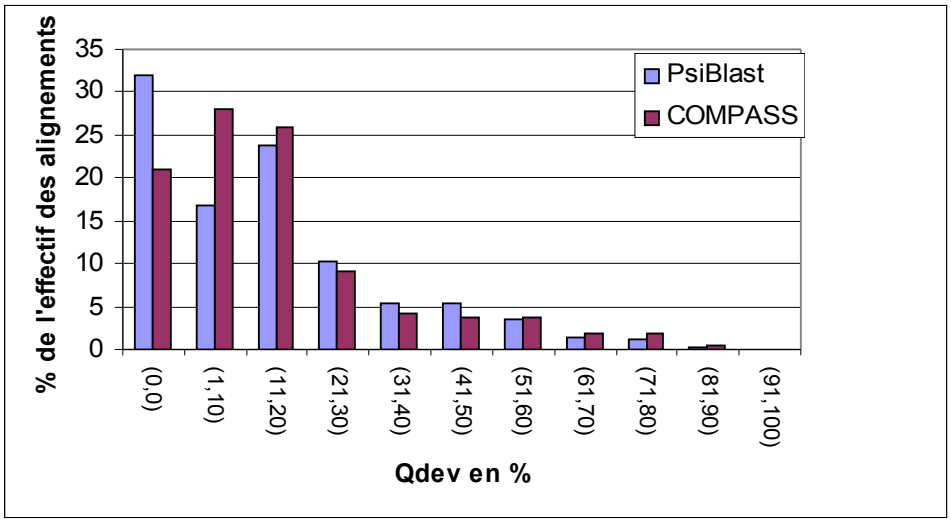


Figure 48 : Distributions des Qdev pour les alignements réalisés par le logiciel COMPASS (rouge) et PSI-BLAST (bleu) à partir des séquences du signal non significatif du logiciel PSI-BLAST. En abscisse, figurent les intervalles de Qdev (de 1% a 100% par pas de 10%). En ordonnée figure le nombre d'alignements associés à un intervalle de Qdev donné.

Pour évaluer si l'augmentation de qualité mise en évidence par le Qmod est associée à un allongement des régions alignées, les valeurs du paramètre Qdev ont été calculées et reportées Figure 48. Sur cette figure, les distributions des valeurs de Qdev obtenues avec PSI-BLAST et COMPASS sont comparées (Figure 48). La différence majeure entre les deux

distributions est concentrée sur les valeurs de Qdev comprises entre 0 et 10 %. Il apparait que les 10 % d'alignements ne possédant plus un Qdev nul avec COMPASS se retrouvent exclusivement avec des Qdev dans l'intervalle (1%,10%). Ainsi, COMPASS ne favorise pas l'augmentation de la taille des régions alignées, et ce bien que les fragments extraits du signal non significatif soient allongés de 30 acides aminés de part et d'autre de la région alignée. La présence caractéristique de grandes insertions ou délétions entre les séquences de protéines très divergentes est une propriété qui peut expliquer cette incapacité de COMPASS à augmenter la taille des régions alignées.

I.1.1.19 Evaluation de la qualité des alignements des séquences homologues calculés par le logiciel HHsearch.

Une étude similaire a été réalisée sur le logiciel HHsearch avec et sans contribution des scores de prédictions de structures secondaires. Dans la suite de l'analyse, les approches utilisant le logiciel HHsearch avec ou sans les prédictions de structures secondaires seront respectivement définie par les termes HHss ou HHnoss. Comme précédemment, les Qmod des alignements de séquences homologues lointains au sein du signal non significatif obtenus avec les logiciels PSI-BLAST et HHsearch ont été reportés Figure 49 (avec et sans la contribution des prédictions de structure secondaire).

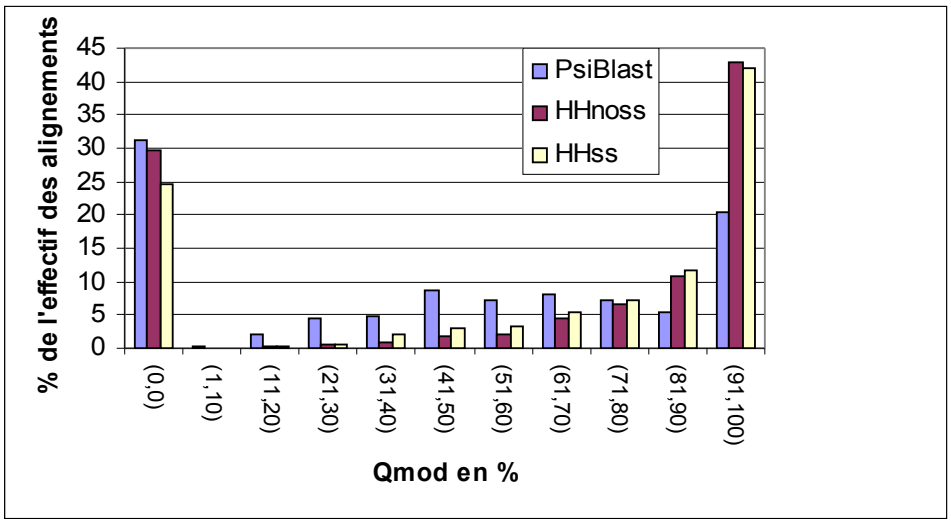


Figure 49 : Distributions des Qmod pour les alignements réalisés par le logiciel HHsearch (HHss, HHnoss) et PSI-Blast à partir des séquences du signal non significatif du logiciel PSI-Blast. En abscisse figurent les intervalles de Qmod. En ordonnée figure le nombre d'alignements associés à un Qmod donné. La distribution des Qmod des alignements proposés par le logiciel PsiBlast est représentée en bleu. Les résultats de l'analyse réalisée par le logiciel HHss sans prédiction de structure secondaire sont présentés en bordeaux. Les résultats de l'analyse réalisée par le logiciel HHss avec prédiction de structure secondaire sont présentés en jaune.

Les distributions observées sont sensiblement différentes de celles obtenues avec le logiciel COMPASS. Tout d'abord, on n'observe pas de diminution importante du nombre d'alignements possédant des Qmod nuls entre PSI-BLAST et HHsearch. L'utilisation des structures secondaires permet néanmoins de rattraper 5 % des cas (contre 10 % avec COMPASS). Par ailleurs, on note une amélioration importante de la qualité des alignements avec 40% des alignements calculés par HHsearch qui possèdent un Qmod supérieur à 90 % au lieu de 20 % avec PSI-BLAST. L'ajout du score de prédiction de structure secondaire ne semble pas affecter significativement cette proportion. Les caractéristiques observées pour les trois méthodes de comparaison profil/profil sont synthétisées sur la Figure 50. La figure illustre bien la meilleure capacité de COMPASS à corriger des alignements entièrement erronés (Qmod nuls) et la meilleure capacité de HHsearch à générer des alignements précis. Toutefois, à ce stade de l'analyse, il n'est pas possible de discerner si la différence de précision est influencée par la taille des alignements proposés par les deux méthodes. En effet ces résultats pourraient résulter d'alignements plus courts du logiciel HHsearch présentant une précision plus importante à l'opposé d'alignements COMPASS de tailles plus importantes associés à un plus grand nombre d'erreur.

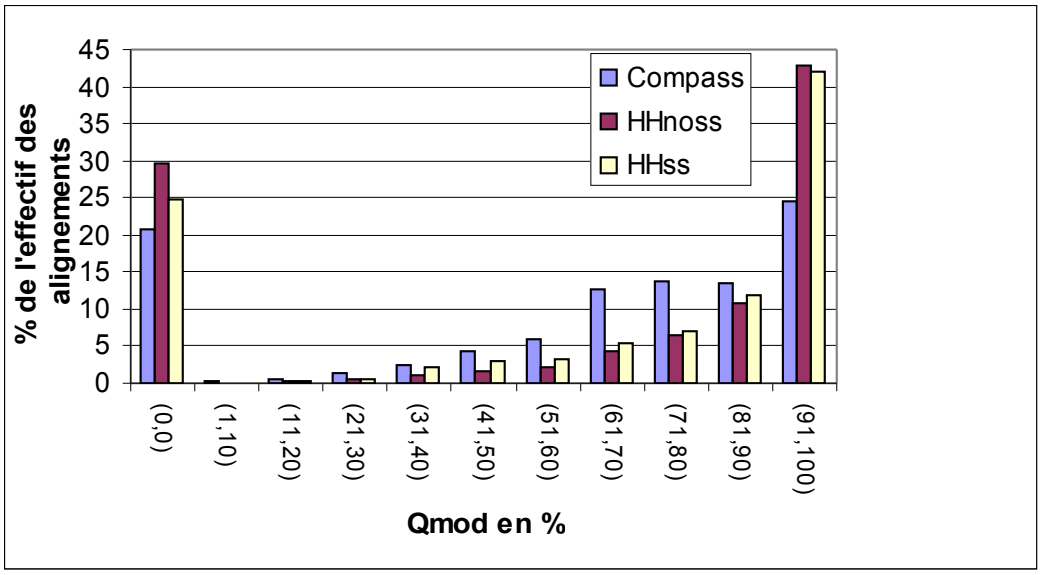


Figure 50 : Distributions des Qmod pour les alignements réalisés par le logiciel HHsearch (HHss, HHnoss) et COMPASS à partir des séquences du signal non significatif du logiciel PSI-Blast. En abscisse figurent les intervalles de Qmod. En ordonnée figure le nombre d'alignements associés à un Qmod donné. La distribution des Qmod des alignements proposés par le logiciel COMPASS est représentée en bleu. Les résultats de l'analyse réalisée par le logiciel HHss sans prédiction de structures secondaires sont présentés en bordeaux. Les résultats de l'analyse réalisée par le logiciel HHss avec prédiction de structures secondaires sont présentés en jaune.

Pour répondre à cette question, les Qdev des alignements obtenus avec le logiciel HHsearch pour les séquences homologues au sein du signal non significatif ont été comparés aux Qdev des alignements obtenus avec les logiciels COMPASS et PSI-BLAST (Figure 51).

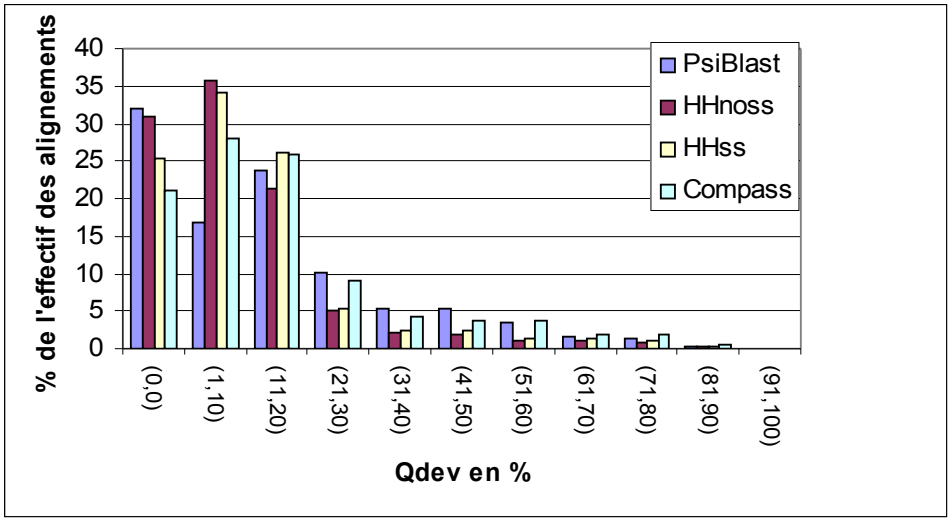


Figure 51 : Distributions des Qdev pour les alignements réalisés par les logiciels PSI-BLAST, HHsearch (HHss, HHnoss) et COMPASS à partir des séquences du signal non significatif du logiciel PSI-BLAST. En abscisse figurent les intervalles de Qdev (de 1 a 100% par pas de 10%). En ordonnée figure le nombre d'alignements associés à un Qdev donné. La distribution des Qdev des alignements proposés par le logiciel PSI-BLAST est représentée en bleu. Les résultats de l'analyse réalisée par le logiciel HHss sans prédiction de structures secondaires sont présentés en bordeaux. Les résultats de l'analyse réalisée par le logiciel HHss avec prédiction de structures secondaires sont présentés en jaune. Les résultats de l'analyse réalisée par le logiciel COMPASS avec prédiction de structure secondaire sont présentés en bleu clair.

Les distributions obtenues avec HHss ou HHnoss sont relativement équivalentes. Pour des valeurs de Qdev comprises entre 1% et 10%, on observe une différence notable entre les méthodes HHsearch et la méthode PSI-BLAST. Les deux variantes de HHsearch génèrent deux fois plus d'alignements (35 % des alignements) possédant ces faibles valeurs de Qdev que PSI-BLAST (15 % des alignements). Cette variation ne s'explique pas par l'amélioration des alignements possédant des Qdev nuls avec la méthode PSI-BLAST car les pourcentages d'alignements possédant des Qdev nuls sont équivalents pour les HHnoss et PSI-BLAST (~30% des alignements). L'interprétation plus probable est que des alignements relativement longs, mais en partie faux, obtenus avec PSI-BLAST ont été raccourcis par les méthodes HHsearch en optimisant la qualité des alignements. Ainsi, l'amélioration de la qualité des alignements par les méthodes HHsearch révélée par l'analyse du Qmod (Figure 49 et Figure 50) se fait au détriment de la longueur des alignements. Cette situation diffère donc sensiblement des résultats obtenus avec le logiciel COMPASS (Figure 48). Par rapport à

COMPASS, la proportion d'alignement avec des Qdev inférieurs à 10 % pour les méthodes Hss et HHnoss montrent que ces dernières génèrent davantage d'alignements de très courtes tailles. Au contraire, pour des Qdev supérieurs à 20 %, la proportion d'alignements COMPASS est 10% supérieure à celles des alignements réalisés avec le logiciel HHsearch. On peut donc résumer les caractéristiques des alignements obtenus avec les séquences du signal non significatifs en trois points :

- Le logiciel COMPASS permet d'obtenir des alignements de plus grande taille mais présentant une précision moins importante que celle du logiciel HHsearch
- Le logiciel HHsearch produit des longueurs d'alignement plus courtes que le logiciel COMPASS et parfois même que PSI-BLAST mais d'une précision plus importante que les deux méthodes.
- L'utilisation des prédictions de structure secondaire permet d'améliorer les performances du logiciel HHsearch en terme de longueur d'alignement (le rapprochant des résultats du logiciel COMPASS) tout en maintenant une forte précision.

IV.3.2.Evaluation des capacités discriminantes des logiciels COMPASS et HHsearch dans le traitement du signal non significatif du logiciel PSI-BLAST.

I.1.1.20Evaluation des seuils de filtrage des méthodes profil/profil

Pour poursuivre cette étude nous nous sommes intéressés aux capacités des logiciels COMPASS et HHsearch à détecter des homologues lointains au sein du signal non significatif généré par le logiciel PSI-BLAST. Dans un premier temps, nous avons étudié le rapport sensibilité/spécificité des logiciels COMPASS et HHsearch afin de déterminer une e-value seuil permettant de détecter un maximum de séquences d'homologues lointains au sein du signal non significatif tout en minimisant l'apparition de faux positifs. Afin d'évaluer la sensibilité et la spécificité en fonction de la e-value, une matrice de confusion a été calculée pour les e-value comprises entre (10^{-3} et 1000, par paliers d'un facteur 10). Classiquement les matrices de confusion sont représentées sous la forme suivante :

	Homologues lointains	Non Homologues
Considérés comme homologues par le logiciel COMPASS	Vrais Positifs	Faux Positifs
Considérés comme non homologues par le logiciel COMPASS	Faux Négatifs	Vrais Négatifs

La sensibilité et la spécificité sont ensuite calculées selon les formules suivantes :

Sensibilité= Vrais Positifs/ (Vrais positifs+Faux Négatifs)

Spécificité=Vrais Négatifs/ (Vrais Négatif+Faux Positifs)

La sensibilité renseigne sur la capacité de détecter le maximum d'homologues pour une e-value donnée, la spécificité sur la capacité discriminative de la méthode à identifier les homologues pour une e-value donnée. La sensibilité et la spécificité obtenues pour chaque e-value ont été reportées pour les différentes approches étudiées (Figure 52).

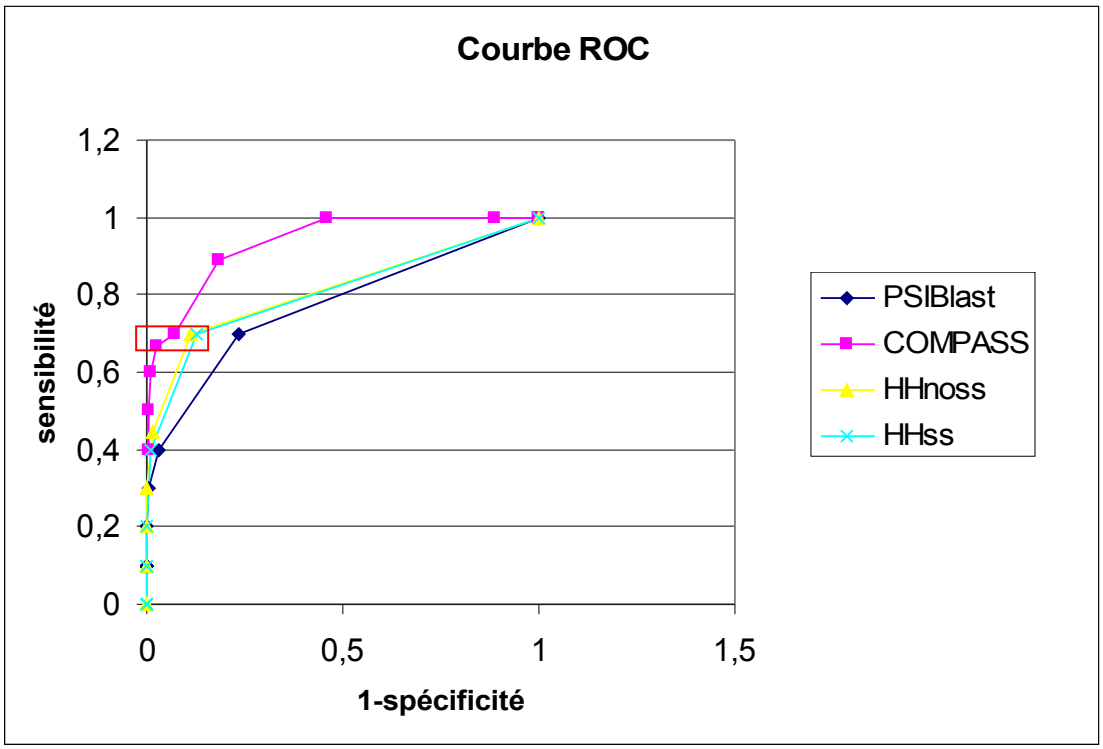


Figure 52 : Comparaison des Courbes ROC du logiciel Psiblast et COMPASS. En abscisse figure la valeur de la spécificité. En ordonnée figure la sensibilité. La courbe ROC associée au logiciel Psi blast est représentée en bleu. En rose figure la courbe ROC associée au logiciel COMPASS. En jaune l'approche HHnoss. En Bleu clair l'approche HHss.

Les 4 courbes ROC suivent des distributions asymptotiques partant de 0 et tendant vers 1. L'aire présente sous la courbe permet d'évaluer quelle approche présente le meilleur rapport sensibilité/spécificité. Nous pouvons ainsi observer que l'approche COMPASS semble présenter les meilleures performances par rapport aux logiciels HHsearch et PSI-BLAST. De plus, on peut observer que le logiciel HHsearch présente lui aussi de meilleures performances que le logiciel PSI-BLAST mais il semblerait que l'utilisation des prédictions de structure secondaire dans l'approche HHss n'aient pas réellement permis une amélioration de la détection des homologues présents au sein du signal non significatif (cette question sera discutée plus en détails à la fin du manuscrit page 157).

Notre objectif est de filtrer un grand nombre de séquences pour réduire l'espace de recherche des homologues lointains. Il nous faut alors trouver le juste équilibre entre sensibilité et spécificité. A partir de la courbe ROC nous avons cherché à déterminer une e-value seuil qui rendent compte de ces contraintes. La zone encadrée en rouge (Figure 52) présente un rapport sensibilité/spécificité raisonnable pour notre objectif. Cette zone correspond à des e-value comprises entre 10^{-4} et 10^{-1} pour le logiciel COMPASS alors qu'elle correspond à des e-value comprise en 50 et 100 pour le logiciel HHsearch. Si on fixait pour HHsearch une e-value seuil à 10^{-3} comme pour COMPASS, seuls 10 % des homologues lointains pourraient être récupérés. Il est intéressant de noter que les e-values des deux approches ne semblent pas s'accorder en termes de faux-positifs attendus (ce point fait l'objet d'une discussion dans le chapitre VI page 157). A partir de ces analyses les e-values sélectionnées comme seuils pour les différentes méthodes sont indiquées dans le Tableau 5.

	COMPASS	HHnoss	HHss
Evalue Seuil	10^{-3}	75	75

Tableau 5 : Tableau récapitulatif des e-values utilisées pour le filtrage du signal non significatif à l'aide des logiciels COMPASS et HHsearch.

Une autre représentation pour comparer les performances des méthodes consiste à suivre l'apparition des faux positifs associée à la détection de nouvelles séquences d'homologues lointains. (courbe de sensibilité des logiciels PSI-BLAST, COMPASS, HHnoss et HHss tracée Figure 53 similaire à celle publié par J. Soeding et présentée en introduction Figure 19 page 53).

Afin d'évaluer les différences de sensibilité entre l'approches HHnoss et HHss nous avons été contraint, pour ces deux approches, de ne pas nous baser sur les e-values calculées par le logiciel HHsearch mais sur le classement des séquences proposé à partir des probabilités. En effet, une information absente de la publication originale de HHsearch mais obtenue sur le site du logiciel HHsearch (http://toolkit.tuebingen.mpg.de/hhpred/help_ov), indique que le calcul des e-values ne tient pas compte de la contribution des prédictions de structure secondaire. Ceci est du au fait que la distribution des scores calculés avec la contribution des prédictions de structures secondaires ne suit pas une loi statistique de la valeur extrême. Nous avons donc utilisé le classement des séquences proposé par le logiciel HHsearch basé sur la probabilité et non sur la e-value. Cette probabilité est calculée en

s'appuyant sur les scores S d'alignement obtenus lors de l'étape de calibration du profil HMM par la relation :

$$proba_{HHsearch} = proba_{Homol}^{score=S} / (proba_{Homol}^{score=S} + proba_{faux-positif}^{score=S})$$

Sur le graphique Figure 53 on note qu'il existe bien une différence de détection entre les méthodes de comparaison profil/profil et le logiciel PSI-BLAST. En effet les méthodes de comparaison de profils semblent minimiser l'apparition des faux positifs parmi le signal non significatif. Parmi les méthodes de comparaison profil/profil employées, le logiciel COMPASS présente les meilleurs résultats. On remarque ainsi que l'apparition des faux positifs reste faible pour 1200 homologues lointains détectés par rapport aux approches PSI-BLAST et HHnoss. Néanmoins l'utilisation des structures secondaires lors de l'approche HHss permet une augmentation notable de la sensibilité.

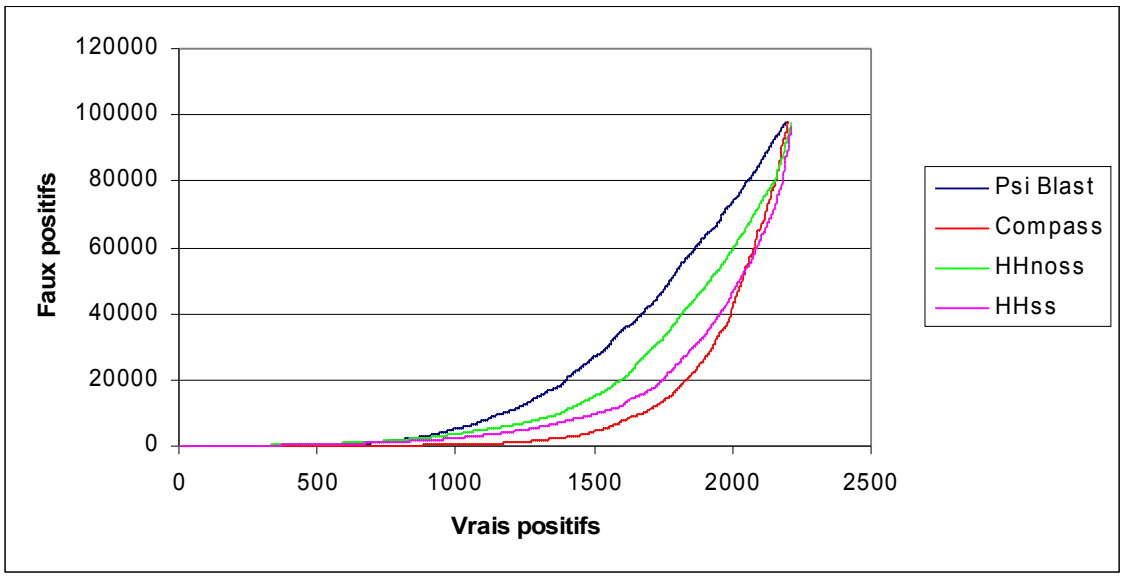


Figure 53 : Représentation de l'apparition des faux positifs en fonction du nombre de vrais positifs détectés par chacune des méthodes. En abscisse figure le nombre de vrais positifs, en ordonnée le nombre de faux positifs. La courbe bleue est associée aux résultats du logiciel PSI-BLAST, la courbe rouge au logiciel COMPASS, la courbe verte au logiciel HHsearch sans l'utilisation des prédictions de structures secondaires, la courbe rose au logiciel HHsearch avec les prédictions de structures secondaires.

I.1.1.21Filtrage avec le logiciel COMPASS

A partir des seuils déterminés par l'analyse globale des faux positifs, la proportion de séquences homologues et non homologues récupérés après filtrage du signal non significatif de PSI-BLAST a été évaluée. Dans un premier temps nous avons analysé les résultats obtenus par le filtrage du signal non significatif par le logiciel COMPASS associé à une e-value seuil de 10^{-3}.

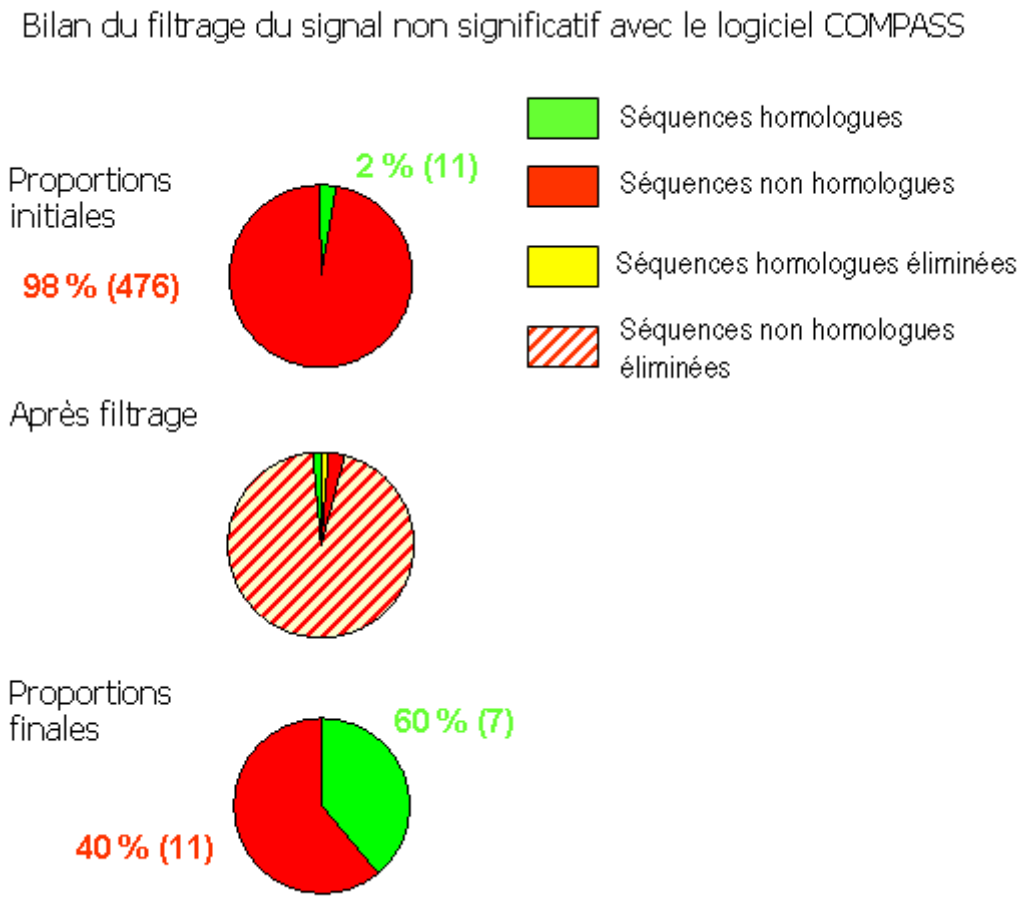


Figure 54 : Diagramme représentant les proportions moyennes de séquences homologues et non homologues filtrées avec COMPASS. En vert, est figurée la proportion de séquences homologues présentes dans le signal non significatif, en rouge, la proportion de séquences non homologues, en hachuré la proportion de séquences non homologues filtrées. En jaune, la proportion de séquences homologues filtrées.

Les résultats du filtrage sont représentés sous forme de diagramme « camembert » (Figure 54[*]). Le diagramme du haut présente la proportion initiale de séquences d'homologues lointains (en vert) et de non homologues (en rouge) au sein du signal non significatif. Cette proportion est relative aux 200 cas étudiés. Le diagramme du milieu permet d'évaluer les quantités de séquences d'homologues lointains et de non homologues conservées et éliminées après filtrage avec la méthode COMPASS. Le diagramme du bas présente la proportion finale de séquences homologues et non-homologues qu'il est possible de sélectionner comme homologues potentiels sur la base d'une e-value seuil de 10^{-3}.

[*] Le chiffre de 476 sur la figure est différent de l'effectif moyen de 600 séquences montré au chapitre III. L'origine de cette différence est due à un problème rencontré lors du traitement des sorties des fichiers PSI-BLAST. Dans la nouvelle version du programme utilisée au chapitre V, ce problème a été corrigé. Par manque de temps, la correction n'a pu être introduite pour cette partie de l'étude. J'espère pouvoir effectuer cette correction avant Noël et vous renvoyer des résultats sur l'ensemble de la base de données dans une version améliorée.

Avant filtrage, la proportion de séquences d'homologues lointains représente 2% des séquences présentes au sein du signal non significatif. Le filtrage avec le logiciel COMPASS permet d'éliminer ~90% des séquences non homologues du signal non significatif et entraine une perte de 36 % des séquences d'homologues lointains présents dans ce signal. Parmi ces séquences homologues perdues 29 % correspondent en fait aux séquences dont le Qmod était nul et dont la détection par PSI-BLAST était dès le départ fortuit. L'approche ne perd donc pas 36 % des séquences potentiellement intéressante mais plutôt 36-29=7% ce qui constitue un taux de perte tout à fait acceptable. Au final après filtration les séquences homologues représentent environ 35% de l'effectif restant.

I.1.1.22Filtrage avec le logiciel HHsearch

De façon similaire, nous avons ensuite analysé les résultats obtenus par la filtration du signal non significatif du logiciel HHsearch associé à une e-value seuil de 75 (Figure 55[*]). De manière identique à la représentation utilisée pour la filtration à l'aide du logiciel COMPASS nous pouvons observer que :

- Les séquences d'homologues lointains représentent initialement 1% des séquences du signal non significatif.
- Le filtrage avec le logiciel HHsearch permet une élimination de 90% des séquences non homologues présentes et une perte de 46% des séquences homologues du signal non significatif. Parmi ces séquences non détectées ~80 % correspondent en fait aux séquences dont le Qmod était nul et dont la détection par PSI-BLAST était dès le départ fortuite. La perte de séquences réellement intéressantes est donc de l'ordre de 9 %.
- Au final, après filtrage, les séquences homologues représentent environ 25 % de l'effectif restant.

[*]

[*] cf note Figure 54

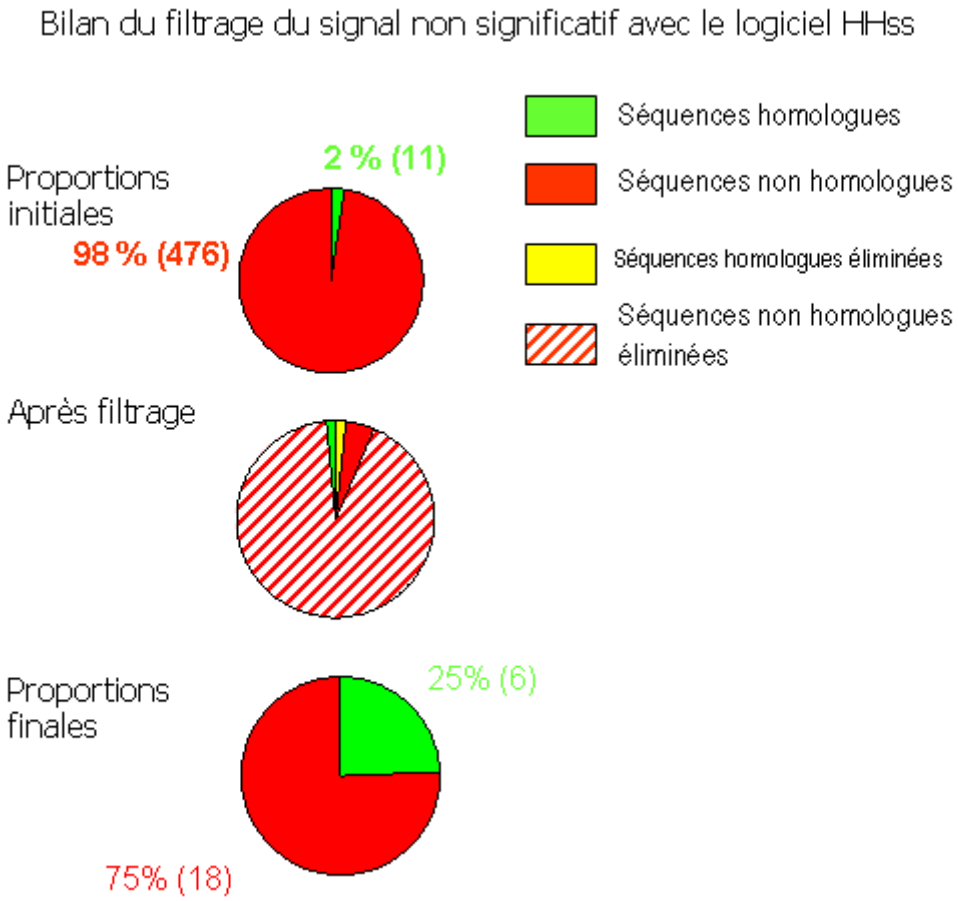


Figure 55 : Diagramme représentant les proportions moyennes de séquences homologues et non homologues filtrées avec la méthode HHss. En vert, est figurée la proportion de séquences homologues présentes dans le signal non significatif, en rouge la proportion de séquences non homologues, en hachuré la proportion de séquences non homologues filtrées. En jaune le pourcentage de séquences homologues filtrées.

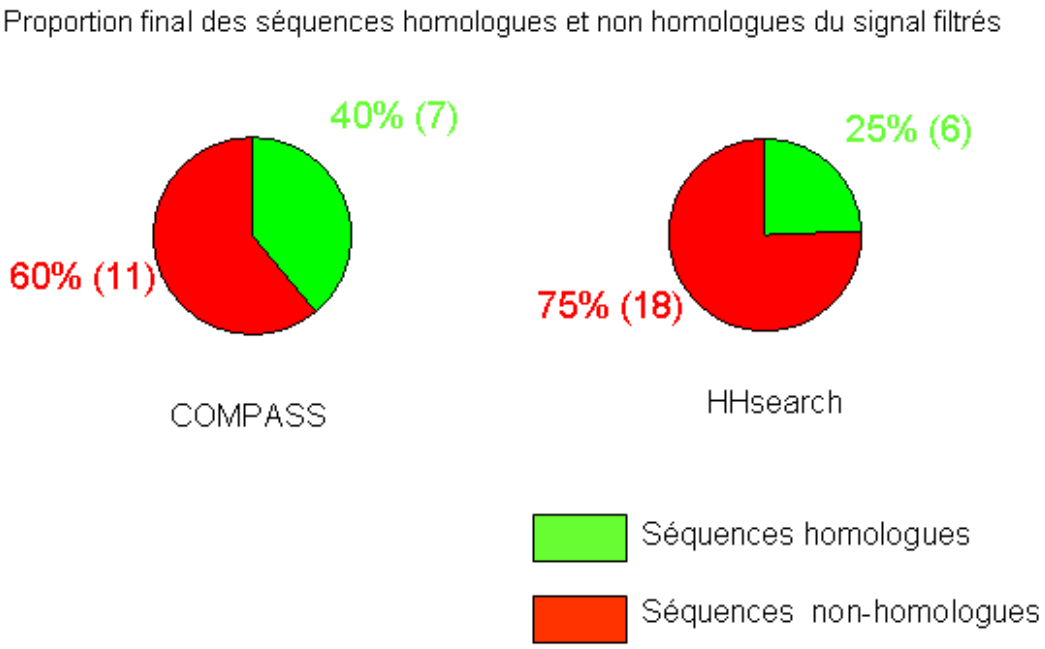


Figure 56 : Récapitulatifs des proportions finales de séquences homologues et non homologues à la suite de la filtration du signal non significatif.

Les analyses obtenues avec HHsearch et COMPASS montrent que les méthodes de comparaison profil/profil permettent, avec des seuils adaptés, d'éliminer parmi le signal non significatif une proportion importante de séquences non homologues tout en limitant les pertes en séquences d'homologues lointains. Après filtrage, les séquences sélectionnées contiennent ainsi ~35% de séquences homologues avec le logiciel COMPASS et ~25% avec le logiciel HHsearch. Nous avons noté pour les deux méthodes une perte d'environ 40 % des homologues lointains lors du filtrage. Néanmoins, ce chiffre contient 80% de séquences dont le Qmod est nul (calculé au chapitre II), c'est-à-dire pour lesquelles l'alignement calculé par PSI-BLAST est fortuit (cf chapitre II Figure 25 page 70). La perte d'homologues lointains qui étaient partiellement bien aligné par PSI-BLAST est donc plutôt de l'ordre de 10 %.

L'analyse, dans la section précédente, de la qualité des alignements (Qmod et Qdev) produits par les deux méthodes COMPASS et HHsearch suggère que leurs modes de reconnaissance de l'homologie lointaine sont relativement distinctes. De plus, les disparités obtenues dans les valeurs seuils de e-value suggèrent que les modes de normalisation des scores diffèrent également sensiblement. Pour ces raisons, nous avons par la suite cherché à évaluer si les deux approches étaient redondantes ou si leurs capacités de détection des homologues lointains pourraient s'avérer complémentaires.

Au début de mon analyse, j'ai estimé que la e-value était le paramètre le mieux adapté pour établir les seuils de filtrage des homologues lointains. Ce n'est qu'au cours de la rédaction de la thèse que je me suis rendu compte que les e-values calculées par HHsearch avec la contribution des structures secondaires (HHss) n'était pas le paramètre optimal. Comme mentionné précédemment, le paramètre de probabilité fournit par le programme est dans ce cas là meilleur (http://toolkit.tuebingen.mpg.de/hhpred/help_ov). Je n'ai pas encore eu le temps d'effectuer l'analyse avec les probabilités. Pour cette raison, dans la suite de l'étude, la capacité de HHsearch à filtrer les homologues lointains pourrait être sous-estimée. J'espère avant la soutenance parvenir à générer de nouveaux résultats qui évaluent la valeur seuil par le paramètre de probabilité et non par le paramètre de e-value.

IV.3.3. Croisement des résultats obtenus avec les logiciels COMPASS et HHsearch. Une méta-approche pour la détection des homologues lointains.

Afin d'améliorer la spécificité de la stratégie de filtrage nous avons envisagé le croisement des résultats entre les méthodes COMPASS et HHsearch. Dans ce but, pour chaque ensemble de séquences sélectionné par une première étape de filtrage, nous avons évalué la capacité de filtrage de la seconde méthode. Cette stratégie de combinaison de deux techniques de comparaison profil/profil est particulièrement rapide puisque l'étape limitante de construction des alignements multiples pour les séquences du signal non significatif a déjà été réalisée à la première étape de filtrage. Comme précédemment nous avons évalué les proportions de séquences homologues et non-homologues éliminées.

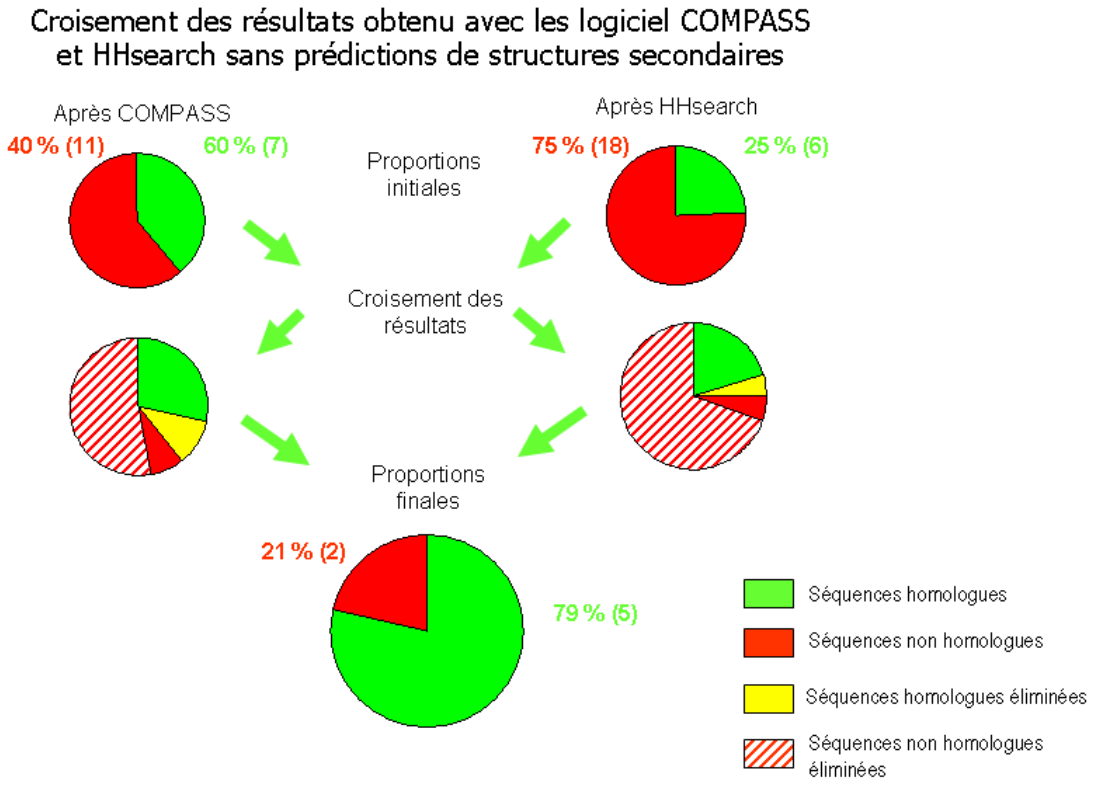


Figure 57 : Diagramme représentant la proportion de séquences homologues et non homologues filtrées. En vert est figurée la proportion de séquences homologues présentes dans le signal non significatif en rouge la proportion de séquences non homologues, en hachuré la proportion de séquences non homologues filtrées. En jaune le pourcentage de séquences homologues filtrées.

Les diagrammes du haut rappellent les proportions moyennes d'homologues lointains et de non homologues dans les séquences sélectionnées après traitement soit avec COMPASS soit avec HHsearch. Les diagrammes du milieu détaillent la quantité de séquences éliminées par le traitement avec la seconde méthode. Enfin, le diagramme du bas représente les proportions d'homologues lointains et de non homologues récupérées après croisement entre les deux méthodes.

Nous pouvons observer que l'application d'une deuxième méthode de comparaison sur les séquences sélectionnées par une première permet de conserver près de 75 % des homologues lointains tout en éliminant 90 % des séquences non-homologues. Au final, la proportion de séquences homologues après croisement est 3,5 fois plus importante que le nombre de séquences non-homologues restantes. Le croisement des résultats obtenus par les deux approches permet une amélioration importante de la spécificité du tri réalisé sur le signal non significatif du logiciel PSI-BLAST tout en limitant la perte de séquences homologues lors du double filtrage.

IV.3.4. Double filtrage avec les prédictions de structures secondaires suivi des comparaisons profil/profil.

Comme nous l'avons mentionné, l'étape limitante en termes de temps de calcul de cette stratégie de filtrage du signal non significatif par des méthodes de comparaison profil/profil se situe lors de la construction des profils. En vue d'optimiser le temps requis pour filtrer le signal non significatif, nous avons envisagé d'utiliser les méthodes de filtrage basées sur les prédictions de structure secondaire développées dans le chapitre III. Ces approches ne présentent pas à elles seules un potentiel discriminant assez fort pour permettre une identification des séquences d'homologues lointains. Toutefois, l'utilisation des prédictions de structures secondaires sur séquences uniques est une étape rapide à effectuer.

Nous avons choisi d'utiliser le paramètre Qsecpred correspondant à une analyse des prédictions de structures secondaires locale (restreintes à la région alignée par PSI-BLAST). L'utilisation des prédictions de structures secondaires prédites sur la séquence globale n'a pas donné de résultats assez convaincants pour être exploitée ici (cf chapitre III).

Comme précédemment nous avons déterminé les proportions de séquences homologues et non homologues du signal non significatif avant et après filtration pour nos 200 cas d'étude.

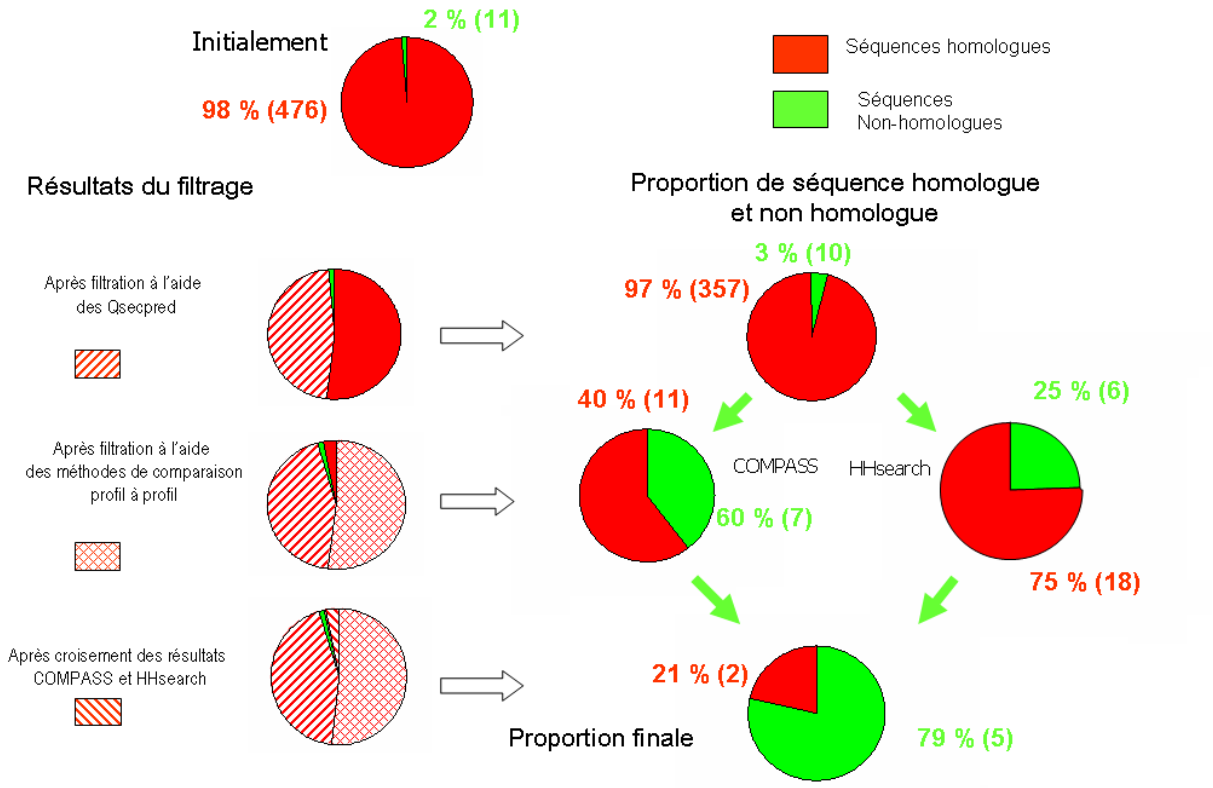


Figure 58 : Diagramme représentant la proportion de séquences homologues et non homologues filtrées pour des valeurs de Qsecpred supérieur à 20 % puis par l'utilisation des logiciels COMPASS (evalue seuil de 0.001) et HHsearch (evalue seuil de 75). En vert est figurée la proportion de séquences homologues présentes dans le signal non significatif, en rouge la proportion de séquence non homologue, les hachures la proportion de séquence non homologue filtrée à l'aide du Qsecpred, les croisillons rouges la proportion de séquence non homologue filtrée à l'aide du logiciel COMPASS et HHsearch.

Sur la gauche sont représentées les proportions de séquences homologues et non-homologues éliminées aux différentes étapes de filtrage. Sur la droite sont représentées les proportions finales de séquence homologues et non-homologues. Comme nous l'avons vu dans le chapitre III, l'utilisation des Qsecpred permet une élimination de 48% des séquences non homologues. En considérant que certains homologues lointains sont a priori « non détectables » (séquences homologues présentant initialement un alignement PSI-BLAST associé à un Qmod nul), la filtration permet de conserver 90% des séquences homologues. Les résultats du filtrage par le croisement des méthodes de comparaison profil/profil permet de sélectionner à peu près la même proportion de séquences que sans l'étape de prédiction de structures secondaires. L'étape de « préfiltrage » avec les prédictions locales de structures secondaires permet de réduire de 48% le nombre de profils à construire.

A titre d'exemple, pour évaluer le gain de temps apporté par ce préfiltrage, nous avons chronométré pour 5 cas d'étude le temps nécessaire au traitement de l'ensemble du signal non significatif. Les temps requis pour filtrer le signal non significatif avec ou sans l'étape de prédictions de structure secondaire ont été comparés (Figure 59). Ces résultats montrent qu'en moyenne le temps de calcul nécessaire à la construction des profils est réduit d'environ 44%.

Séquences	Superfamille	T1=Temps d'exécution sans l'utilisation du Qsecpred (secondes)	T2=Temps d'exécution avec pré-filtrage Qsecpred (secondes)	T1/T2 * 100
d1avgi	b.60.1.3	16358 (~4h30)	8343 (~2h15)	51%
d1b0ua	c.37.1.12	14056 (~4h)	9713 (~2h30)	70%
d1cc5	a.3.1.1	12345 (~3h30)	3983 (~1h)	32%
d1coza	c.26.1.2	18219 (~5h)	11732 (~3h15)	64%
d1ei9a	c.69.1.13	13668 (~3h45)	8059 (~2h15)	58%
Moyenne		14929 (~4h)	8366 (~2h15)	56%

Figure 59 : Tableau récapitulatif du temps nécessaire à l'ensemble des l'analyse avec et sans filtration à l'aide du Qsecpred. Temps de calcul estimé sur un cluster de 10 processeurs (Intel Xeon 3GHz).

IV.3.5. Exemple d'analyse : Deux domaines de liaison au NADP d1hxha et d1dih1 appartenant à la même superfamille.

Pour conclure ce chapitre, nous avons choisi d'illustrer les résultats obtenus au cours du chapitre IV en prolongeant l'exemple développé au cours du chapitre II correspondant aux deux domaines de liaison au NADP : d1hxha et d1dih1

Psi-Blast output

```
> d1dih_1 c.2.1.3 Dihydrodipicolinate reductase [ E. coli ]

 Evalue =  365 Identities = (18%)

Query:  8   VALVTGGASGVGLEVVK-LLLGEGAKV 33
              + G    +G ++++  L  EG ++
Sbjct:  6   RVAIAGAGGRMGRQLIQAALALEGVQL 32
```

Figure 60 Récapitulatif des propriétés de l'alignement PSI-BLAST dans le signal non significatif de la requête effectuée avec le domaine d1hxha.

Nous avons vu au chapitre II (page 76) que le logiciel PSI-BLAST proposait, au sein du signal non significatif, un alignement entre les séquences de ces deux domaines sur une longueur de 26 résidus avec 18% d'identité (Figure 60). Cet alignement est localisé au niveau de trois structures secondaires structuralement superposable et correspond à un Qmod de 70%. L'étape de prédiction de structure secondaire sur les deux séquences montre que leur alignement correspond à un Qsecpred de 57 % largement supérieur au seuil de 20 % (Figure

61). Le Qsecpred obtenu nous permet de sélectionner cet alignement pour une analyse à l'aide des méthodes de comparaison profil-profil. Un profil est construit pour chacune des séquences de l'alignement. Les profils sont ensuite comparés à l'aide du logiciel COMPASS et HHsearch. Le croisement des donnés nous permet de conserver cet alignement après filtrage (Figure 62).

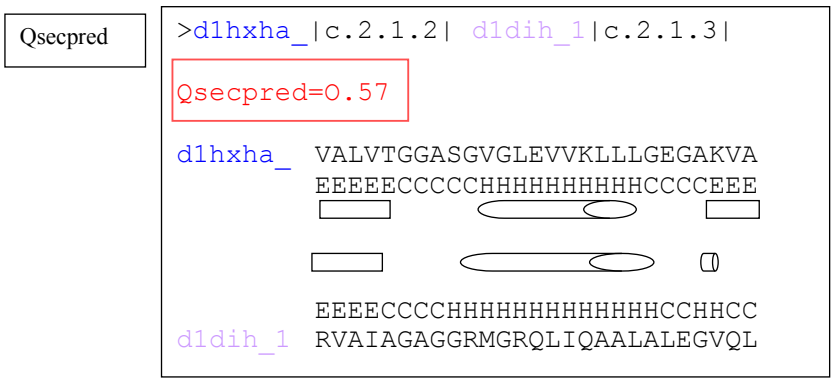


Figure 61 Récapitulatif des propriétés de l'alignement des prédictions de structures secondaires. Les rectangles représentent les brins, les hélices sont représentées par des cylindres.

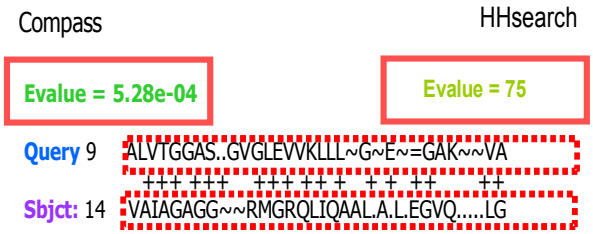


Figure 62: Récapitulatif des scores obtenu lors du croisement des méthodes de comparaison profil à profil COMPASS ET HHsearch.

D'après les valeurs seuils déterminées au cours de ce chapitre nous pouvons voir que, selon nos critères, les deux scores sont significatifs et nous permettent de suggérer la relation d'homologie existant entre les séquences d1hxha et d1dih. Toutefois il est intéressant de noter que si le score du logiciel COMPASS correspond aux valeurs considérées comme significatives, la e-value calculée par le logiciel HHsearch, ne permettait pas de présumer de la relation d'homologie lointaine et ce malgré un alignement associé à une réalité structurale.

Afin de mieux comprendre pourquoi cette région est suffisante pour établir une relation d'homologie lointaine, nous avons étudié les spécificités fonctionnelles et structurales de chaque domaine d1hxha et d1dih1. Ces deux domaines sont impliquées dans des réactions enzymatique de type oxydo/reduction mais sur des substrats bien différents : plusieurs types

de composés stéroïdiques pour 1hxh, et le dyhydrodipicolinate (DHPR) pour le domaine didh. Toutefois ces réactions enzymatiques ont en commun l'utilisation d'un cofacteur : le NAD.

L'analyse de la portion de séquence détectée du domaine 1dih montre que ce domaine porte l'empreinte d'un motif consensus « (V/I)(A/G)(V/I)XGXXGXXG » conservé chez un grand nombre de NADP déshydrogénase et utilisé pour identifier le repliement responsable de la liaison au NAD. Ce motif est localisé sur la boucle entre le premier brin et la première hélice du domaine d1dih (Figure 63). La portion de séquence détectée au niveau du domaine d1hxh ne contient que partiellement le motif consensus. Toutefois la nature des acides aminés reste très proche de celle observée pour le domaine d1dih et des études montrent que cette boucle était elle aussi impliquée dans la liaison au NAD . Si initialement ces séquences sont très divergentes, la présence de cette zone conservée est ici le résultat d'une contrainte fonctionnelle commune aux deux protéines.

Ce cas pratique illustre ainsi l'intérêt qu'il peut exister dans l'étude de certains alignements de petites tailles détectées de manière non-significative. En effet la détection de ces régions présentant une conservation locale plus importante peut être un indicateur de l'existence d'un repliement commun entre deux séquences très divergentes.

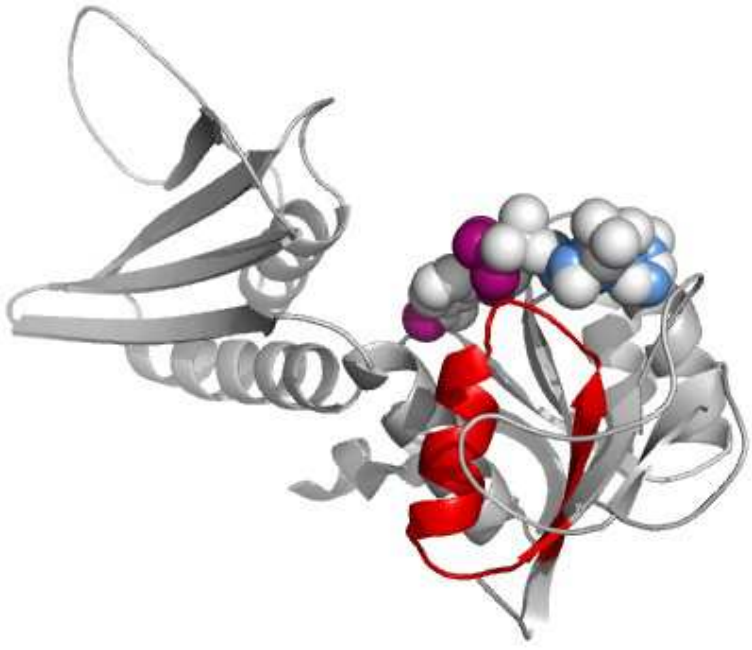


Figure 63 : Représentation du domaine d1dih, lors de sa liaison au NADP. Le domaine d1dih est représenté dans un type « cartoon ».La portion de séquence alignée entre la protéine d1hxh et d1idh est représentée en rouge. Le NADP est figuré selon une représentation type « sphères de Van der Wals » La figure est réalisée à l'aide du logiciel Pymol(version 0.99).

IV.4.Conclusion

Ce chapitre nous montre que les méthodes de comparaison profil-profil fournissent une approche intéressante pour le traitement du signal non significatif. Elles permettent une amélioration de la qualité locale de l'alignement mais n'améliore pas le taux de recouvrement des alignements initiaux proposé par le logiciel PSI-Blast. Le logiciel COMPASS propose des alignements de plus grandes tailles mais parfois de moindre qualité que les alignements plus courts proposés par le logiciel HHsearch (avec et sans prédictions de structures secondaires). Au niveau de la détection, ces approches permettent une élimination d'environ 90% des séquences non homologues présentes parmi le signal non significatif. Environ 40% des homologues ne sont pas détectés, mais il s'avère qu'environ 80% des homologues non détectées présentent un alignement fortuit parmi le signal non significatif. Sur un plan comparatif, le logiciel COMPASS se révéle plus efficace que le logiciel HHsearch lors de la recherche d'homologues lointains malgré une littérature en faveur du logiciel HHsearch. Cependant, lors de l'utilisation du logiciel HHsearch, l'utilisation des structures secondaires permet une amélioration de la détection des homologues. De plus nous avons montré que ces deux approches sont complémentaires et qu'un croisement des résultats obtenus par ces deux approches permet d'augmenter la spécificité des résultats obtenus indépendamment.

Une des limites liée à l'utilisation des méthodes de comparaison profil-profil est que la construction des profils nécessite des ressources informatiques importantes, notamment dans le cas d'une étude à grande échelle. Dans ce cadre, nous avons observé que l'utilisation des comparaisons de prédictions de structure secondaire d'un point vue local par le calcul du Qsecpred permettait de diminuer d'environ 45% le nombre de profils nécessaires à l'étude du signal significatif accélérant ainsi le temps nécessaire à la recherche d'homologues lointains.

A partir de ces résultats, nous avons envisagé de mener une étude systématique de 100 protéines issues de la signalisation des dommages de l'ADN. Au cours du chapitre 5, une partie des résultats obtenus sera présentée ainsi qu'une partie des travaux ayant conduit à la publication de deux articles.

CHAPITRE IV : Filtrage du signal non significatif à l'aide des méthodes de comparaison profil/profil.

Chapitre V :Applications

Analyse des protéines impliquées dans la signalisation des dommages de l'ADN chez la levure

V.1.Introduction

Des interactions multiples et transitoires entre protéines coordonnent la progression ordonnée des évènements intervenant autour des molécules d'ADN. A partir de l'observation de la succession rapide de partenaires sur une même molécule d'ADN, ont été développés les concepts de modularité des protéines intervenant sur l'ADN, ainsi que de versatilité des repliements qui leur sont associés .

Les protéines impliquées dans la signalisation et la réparation des dommages de l'ADN sont généralement des protéines de grande taille organisées en plusieurs domaines, chacun de ces domaines pouvant présenter des fonctions et interactions spécifiques. Cette modularité facilite la coordination des fonctions biochimiques portées par les différents domaines, sans contraindre ces domaines à adopter une position fixe les uns par rapport aux autres. L'attachement flexible des domaines au sein des protéines modulaires permet le couplage entre plusieurs évènements biochimiques indépendants. Le corollaire de cette modularité est l'utilisation de plusieurs points de contact lors de l'interaction d'une ou plusieurs paires de protéines entre elles. Ceci facilite la succession rapide d'interactions dans des situations où le réarrangement de complexes est plus fréquent que la dissociation complète de ces complexes. De plus, c'est un moyen d'obtenir une haute affinité à partir d'interactions individuelles faibles. Enfin, l'utilisation de plusieurs points de contact permet à l'interaction d'être régulée par de multiples voies.

Une autre caractéristique des protéines impliquées dans la signalisation et la réparation des dommages de l'ADN est leur utilisation répétée d'un petit nombre de domaines structuraux ou modules. De manière remarquable, ces domaines peuvent interagir exclusivement avec des protéines dans un certain contexte biologique, mais interagiront avec des protéines et de l'ADN simultanément dans un contexte différent.

La compréhension du rôle des protéines impliquées dans la signalisation et la réparation des dommages de l'ADN passe par la détection de leurs différents domaines, leur délimitation exacte et la résolution de leur structure par RMN, cristallographie et modélisation. Or la détection des domaines de protéines pose plusieurs problèmes.

Tout d'abord, on ne dispose à ce jour que de très peu de séquences pour lesquelles la totalité de l'organisation en domaines a pu être détectée. En effet, comme il a été décrit dans ce manuscrit, il peut exister des divergences importantes de séquences entre deux domaines de même famille. De ce fait, l'identification de domaines repose sur la détection d'homologues lointains.

De plus, la délimitation exacte des domaines reste délicate. Dans le cadre d'une analyse structurale, la présence (ou absence) d'un élément de structure secondaire associé au domaine étudié peut entraîner l'impossibilité d'obtenir un peptide structuré. Or, l'évolution d'un domaine peut conduire à l'apparition ou la disparition d'éléments de structure secondaire et les profils disponibles dans les bases de données ne permettent pas toujours de détecter ces éléments supplémentaires de manière correcte.

De nombreux travaux sur les protéines de la signalisation et la réparation des dommages de l'ADN sont réalisés chez la levure, organisme modèle de la cellule eucaryote. Toutefois, les distances phylogénétiques entre la levure et les organismes supérieurs restent une limite à l'exploitation de ces résultats. Pour de nombreuses protéines, la divergence de séquences entre l'homme et la levure ne permet pas de transposer les résultats observés d'un organisme à l'autre.

Nous avons choisi d'appliquer notre approche de filtrage du signal non significatif du logiciel PSI-BLAST à grande échelle, afin de permettre la détection de nouvelles cibles d'étude pour une analyse structurale au sein des protéines de la signalisation et la réparation des dommages de l'ADN de la levure, et afin d'identifier les éventuelles orthologies existant entre des protéines humaines et des protéines de la levure. Ce chapitre décrit la stratégie employée lors de cette analyse à grande échelle, et illustre les résultats obtenus en s'appuyant sur trois exemples : l'étude de fragments issus des protéines de levure RAD9, XRS2 et NEJ1.

V.2.Méthodes

V.2.1. Sélection des protéines d'intérêt.

A partir de la classification fonctionnelle fournie par la banque SGD (http://www.yeastgenome.org/), l'ensemble des séquences de protéines de *Saccharomyces cerevisiae* impliquées dans la signalisation des dommages de l'ADN ont été récupérées. D'autres protéines potentiellement intéressantes dans le contexte de la réparation (associées au remodelage de la chromatine et à la formation du kinétochore) ont également été suggérées par l'un de notre collaborateur généticien, Carl Mann. Un ensemble de 100 protéines d'intérêt a ainsi pu être sélectionné.

Pour chaque séquence sélectionnée, un premier alignement a été généré à partir des séquences d'espèces proches de *S. cerevisiae*, généralement absentes de la base de données nr mais disponibles directement sur le serveur SGD . Ces espèces sont *S. bayanus, S. mikatae, S paradoxus, S. kudriavzevii*, *S. castellii* et *S. kluyveri*. Les séquences de l'alignement possédaient rarement des identités de séquences inférieures à 40 %. Elles ont été alignées avec le logiciel Clustalw, puis le profil correspondant a été utilisé pour effectuer la première itération de recherche PSI-BLAST. Les expertises manuelles que nous avons réalisées ont montré que dans de nombreux cas, cette procédure d'enrichissement initial augmentait les chances de détecter des homologues lointains pour des séquences fortement divergentes.

V.2.2.Isolement des régions structurées.

A l'aide du logiciel DISOPRED (version 2.1) , une prédiction de régions déstructurées a été réalisée sur les 100 séquences de protéines impliquées dans la signalisation des dommages de l'ADN. Nous avons alors défini les zones structurées comme les portions de séquences séparées du reste de la protéine par des régions désordonnées plus longues que 30 acides aminés. Les délimitations de chacune de ces zones structurées ont été répertoriées et comparées aux délimitations des domaines répertoriés dans la banque de domaines Pfam (octobre 2006) (ftp.sanger.ac.uk). Seules les portions de plus de 30 acides aminés ont été considérés dans la suite. Dans un premier temps, les portions de séquences prédites comme structurées mais contenant un ou plusieurs domaines connus ont été exclues de la suite de l'analyse. En effet, dans le cadre de mon travail de thèse, je me suis focalisé uniquement sur les portions prédites comme structurées et ne contenant aucun domaine connu afin d'explorer le potentiel de détection de nouveaux domaines par l'approche que j'ai développée.

V.2.3.Recherche de séquences homologues.

Pour effectuer les recherches d'homologie lointaine sur les portions structurées présélectionnées, une banque de séquences spécifiquement eucaryotes a été construite à partir de la banque swissprot-Trembl. Plus de 50% des séquences de cette banque appartiennent aux bactéries, virus et archaebactéries qui ne nous intéressent pas dans un premier temps. L'utilisation d'une banque filtrée eucaryote, dont les séquences bactériennes, virales et archaebactérienne ont été éliminées, permet de réduire les risques d'inclure dans les profils des séquences non homologues et de provoquer ainsi la divergence du programme PSI-BLAST. Dans la suite de ce chapitre, la banque de séquences eucaryotes sera nommée « sptr_euk ».

Une recherche d'homologues a été réalisée sur cette banque avec le logiciel PSI-BLAST (4 itérations, e-value d'inclusion égale à 10^{-4}). Pour chaque portion étudiée, les séquences du signal significatif (e-value inférieure à 10^{-3}) ont été utilisées pour construire un profil dit de référence. Chaque séquence du signal non significatif (e-value comprise entre 10^{-3} et 1000) a ensuite été analysée afin de détecter des homologues lointains liés à la séquence étudiée. Pour cela, de manière similaire à l'approche présentée au cours des chapitres III et IV, la portion de séquence alignée dans le signal non-significatif a été élargie de 50 résidus de part et d'autre. Les séquences élargies ont été nommées « séquences cibles »

V.2.4.Filtration du signal non significatif à l'aide des prédictions de structures secondaires.

Sur chaque portion étudiée dans le profil de référence, une prédiction de structure secondaire a été réalisée à l'aide du logiciel PSIPRED (version 2.45) à partir d'une recherche d'homologues réalisée sur la banque sptr_euk. Sur chaque séquence cible élargie du signal non-significatif, une prédiction de structure secondaire a été réalisée sur séquence unique à l'aide du logiciel PSIPRED.

A partir des prédictions de structures secondaires effectuées sur les séquences références et sur chaque séquence de leur signal non-significatif respectif, le paramètre Qsecpred, caractérisant la qualité de l'alignement local des structures secondaires, a été calculé (chapitre III). Les alignements du signal non significatif présentant un Qsecpred inférieur à 20 % ont été exclus de la suite de l'analyse. Pour les alignements présentant un Qsecpred supérieur à

20 %, la séquence associée à l'alignement a été sélectionnée pour une analyse à l'aide des méthodes de comparaison profil/profil.

V.2.5.Construction des profils.

A partir des séquences détectées de manière significative suite aux itérations PSI-BLAST, les alignements par paire de séquences ont été convertis en alignements multiples et constitue les profiles référence de notre étude. Pour cela, selon une démarche similaire à celle effectuée par le logiciel PSI-BLAST lors de la construction des matrices de score position spécifique (PSSM), aucune insertion n'est conservée au niveau de la séquence référence et les résidus des séquences homologues responsables des insertions ont été éliminés.

Pour chaque séquence du signal non-significatif, une recherche d'homologues sur la banque sptr_euk a été effectuée (4 itérations, evalue d'inclusion égale à 10^{-4}). Selon le même principe qu'au V.2.3, les séquences homologues détectées (e-value inférieure à 10^{-3}) ont été alignées. A partir de ces alignements, selon les méthodes employées au cours du chapitre III, des profils exploitables par les logiciels COMPASS et HHsearch incluant les prédictions de structures secondaires ont été construits.

L'ensemble des étapes nécessaires à la constuction des profils est réalisé à l'aide d'un cluster de 10 Processeurs Intel Xeon 3GHz.

V.2.6.Filtrage du signal non significatif à l'aide des méthodes de comparaison profil/profil.

La comparaison des profils références avec chacun des profils des séquences du signal non-significatif a ensuite été réalisée à l'aide des logiciels COMPASS et HHsearch. Deux profils associés à une e-value inférieure à 10^{-3} pour COMPASS et 75 pour l'approche HHss sont considérés comme significativement homologues (Cf. Chapitre IV). Le reste des séquences est exclu de l'analyse.

V.2.7.Prédiction des domaines

L'organisation en domaines des séquences du signal non significatif sélectionnées a été obtenue à partir de la banque Pfam et peut permettre à l'utilisateur d'évaluer si la présence d'autres domaines dans la séquence cible renforce ou infirme la possibilité d'une relation d'homologie lointaine. Au final, un fichier de sortie notifiant l'ensemble des informations associées aux séquences cible a été créé en suivant le format décrit ci-dessous à partir d'un exemple.

```
############################
P06701@2.info     Code de la protéine d'intérêt@ index du domaine
Sir3              Nom de la protéine d'intérêt
  "Silencing protein that interacts with Sir2p and Sir4p, and histone H3 and H4 tails, to es-
  tablish a transcriptionally silent chromatin state; required for spreading of silenced chro-
  matin; recruited to chromatin through interaction with Rap1p"        YLR442C
                  Annotation SGD
############################
Lecture des informations :
>NomDomaine | CodeProtDetectée | EvaluePsiBlast | EvalueCOMPASS (seuil evalue=10-
  3) | EvalueHHsearch (seuil evalue=75) | definition | DelimitationsMatchDansSeqRéf |
  NbSeqProfilRéf | NbSeqProfilCible | DelimitationsMatchDansSeqCible
*NomDuDomainePFAM|DélimitationDuDomaine|
  X,SiLeDomaineRecouvreLaRégionSélectionnée

Exemple :
>P06701@2|P70044|3.01318|1.47e-128|1e-10|Cell division control protein 6. [Xenopus
  laevis (African clawed frog)]|(576, 817)|26|467|('192', '396')
*AAA|191-395|X
```

V.3.Résultats

D'après l'analyse des 100 protéines d'intérêt, 259 portions de séquence prédites comme structure ont été détectées par le programme. Ces portions ont été filtrées de manière à éliminer les séquences présentant un domaine identifié dans la banque Pfam A, puis de manière à ne conserver que les portions de taille supérieure à 30 résidus. Ce filtrage nous a permis d'identifier 70 portions d'intérêt. Ces portions correspondent à des îlots de régions structurées séparés du reste de la protéine par des régions désordonnées de 30 résidus.

Nous avons alors analysé le signal non significatif de PSI-BLAST sélectionné par notre approche pour les 70 portions de séquences. Quatre catégories de cas sont apparues (Figure 64) :

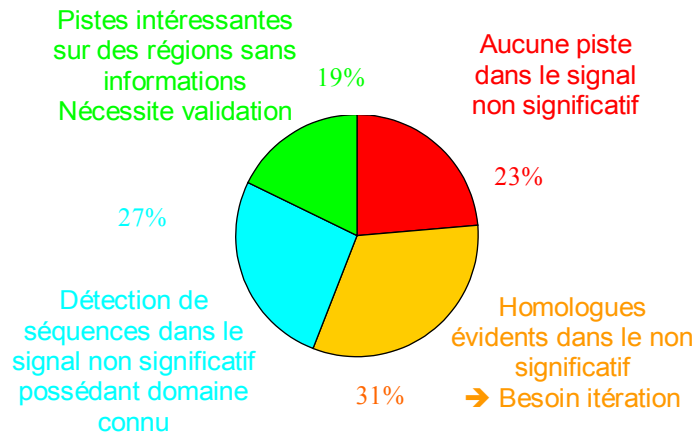


Figure 64 : Répartition des cas de figure obtenus suite à l'analyse des 70 portions de séquences.

- Dans 16 cas (23%, en rouge), les séquences du bruit de fond sont peu nombreuses et ne renvoient, d'après l'annotation des résultats, à aucune analogie structurale ou fonctionnelle avec la séquence d'intérêt. Notre première conclusion est que le signal non significatif est principalement composé de faux positifs.
- Dans 22 cas (31%, en orange), le bruit de fond contient soit des fragments probablement faux positifs comme précédemment, soit des portions de séquences appartenant sans ambiguïté à des homologues ou à des séquences de même architecture que la séquence d'intérêt. Il manque alors à notre approche la possibilité

d'intégrer ces homologues dans le profil de référence afin d'itérer la procédure et de calculer un signal non significatif enrichi.

- Dans 18 cas (27%, en cyan), les portions de séquences sélectionnées dans le signal non significatif correspondent à des domaines SMART, PROSITE, PRODOM, PRINTS (banques de motifs, profils, HMM) qui parfois étaient déjà détectés dans certaines séquences du signal significatif alors que PFAM ne détectait pas le domaine dans la séquence initiale.
- Dans 12 cas (19%, en vert), certaines séquences du signal non significatif pourraient être des homologues lointains potentiels (profils associés riches en diversité, e-values de comparaisons profil/profil largement inférieures aux seuils, contexte fonctionnel cohérent). Ces cas de figure paraissent intéressants à analyser manuellement de façon plus approfondie. Il faudra alors rechercher (i) si il existe déjà un lien dans la littérature entre la portion de séquence d'intérêt et le nouvel homologue potentiel, (ii) si l'on retrouve des motifs fonctionnels communs aux deux séquences, (iii) si l'on peut allonger le court fragment aligné et optimiser cet alignement (cf chapitre 4), (iv) si l'on peut construire un modèle structural de la séquence d'intérêt et le valider, soit en le confrontant à la littérature, soit par l'expérience.

Notre analyse des 70 séquences cibles permet aujourd'hui d'obtenir des informations potentiellement intéressantes et initialement non détectées par la base de donnée PFAM pour 27+19=36% des séquences. Le fait que dans 27% des cas la procédure mise au point permette de retrouver, dans les homologues lointains, des domaines déjà connus et probablement effectivement présent dans notre séquence d'intérêt constitue en quelque sorte une validation de l'approche. Par exemple, dans le cas de Rad1, impliqué dans la réparation par excision de nucléotides (NER), les résultats obtenus soulignent une ressemblance avec un domaine hélicase clairement identifié dans les protéines de la famille Rad54 (hélicase impliquée dans la recombinaison homologue). La proximité entre ces fonctions avait déjà été décrite dans la littérature sur la base d'analyses bioinformatiques ciblées et de résultats expérimentaux de biologie structurale .

Parmi les cibles intéressantes de cette analyses se trouvent aussi des exemples pour lesquels une analyse bioinformatique et/ou structurale avait été conduite au laboratoire et avait permis d'identifier des relations d'homologies lointaines. J'ai personnellement participé

à deux de ces études, l'une durant laquelle Isabelle Callebaut (LMCP, Jussieu) a détecté un domaine TUDOR dans Rad9 de levure et son homologue humain potentiel 53BP1 (Charier G., Thèse de doctorat, 2005) et l'autre durant laquelle j'ai mis en évidence l'existence d'un tandem de domaines BRCT dans Xrs2 de levure et son homologue humain Nbs1 . Le cas de Nej1 de levure a été étudié au laboratoire en collaboration avec Isabelle Callebaut (LMCP, Jussieu ;) avant mon arrivée. Je présenterai maintenant ces trois exemples de manière détaillée, afin de décrire l'apport de la stratégie développée durant ma thèse, et d'identifier les améliorations à apporter pour augmenter l'efficacité de la procédure (les résultats détaillés et annotés de l'analyse peuvent être consultés en annexe page 171).

V.3.1.Analyse d'un domaine de Rad9

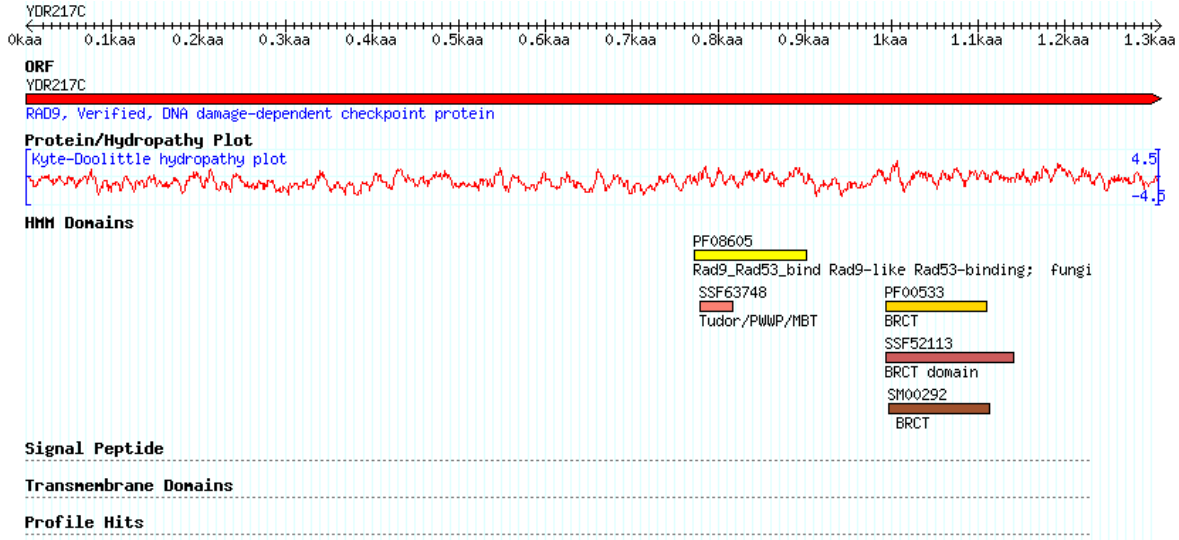


Figure 65 : Représentation schématique des domaines identifiés dans Rad9 à partir de différentes méthodes de détection de domaines (extraites de la base de donnée SGD). Les initiales des codes de domaines indiquent la base de données source utilisée (SSF : Superfamily ; PF : Pfam ; SM :SMART ; PTHR : Panther).

Rad9 est une protéine médiatrice impliquée dans les voies de contrôle de l'intégrité du génome. Elle est nécessaire à l'activation de la kinase Rad53 et à la régulation du cycle cellulaire après dommage de l'ADN. Différentes études suggèrent que Rad9 aurait un rôle dans la reconnaissance de la lésion, la réparation de l'ADN et le maintient de la stabilité génomique (pour revue :). Les deux domaines BRCT C-terminaux de Rad9 sont impliqués dans son oligomérisation . En revanche, le rôle de la région 754 à 947 contenant le fragment récemment annotée par Pfam (Version 21.0) comme impliqué dans la liaison à Rad53 (Rad9_Rad53_bind, en jaune sur la Figure 65) n'est pas connu. Il a été suggéré que cette

région interagirait avec l'histone H3 méthylée sur la lysine 79 . D'un point de vue structural, une étude bioinformatique réalisée par Isabelle Callebaut il y a quelques années avait suggéré que cette portion de Rad9 contiendrait un domaine de type Tudor. Enfin, une étude RMN en cours au laboratoire a montré qu'il existe deux domaines en tonneau β au sein de cette portion de Rad9 (fragments structurés en tonneau β : 762-830 et 834-894), et que le deuxième domaine possède clairement la structure 3D d'un Tudor (N. Lancelot et S. Zinn-Justin, communication personnelle). Nous avons choisi d'utiliser notre approche sur le fragment 754 à 947 de Rad9 afin d'analyser le contenu en vrais et faux positifs du bruit de fond ainsi sélectionné.

L'analyse par PSI-BLAST permet de retrouver 5 séquences de levure homologues à Rad9 et présentes dans la banque PFAM.

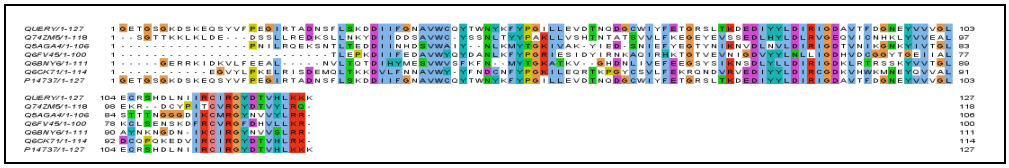


Figure 65 : Alignement de la portion étudiée de Rad9 de Saccharomyces cerevisiae (QUERY) avec les séquences homologues de Ashbya gossypii (Q74ZM5), Candida albicans (Q5AGA4), Candida glabrata (Q6FV45), Debaryomyces hansenii (Q6BNY6) et Kluyveromyces lactis (Q6CK71).

Le signal filtré par notre approche à partir de plus de 1000 séquences comprises dans le signal non significatif de PSI-BLAST comprend 70 séquences. Deux de ces séquences contiennent un domaine Rad9_Rad53_bind (PFAM) et un domaine BRCT (SMART). Ces séquences correspondent à des homologues de Rad9 retrouvés dans les levures *Aspergillus oryzae* et *Aspergillus nidulans*. Leurs e-value étaient initialement de 212 et 11, respectivement. Notre approche les fait ressortir comme homologues, et elles sont aussi retrouvées dans la famille Rad9_Rad53_bind de PFAM.

Une dernière séquence correspondant à Rad9 de *Aspergillus fumigatus* appartient à la famille PFAM et n'est pas retrouvée par notre approche. Cette séquence ne se trouve pas dans le signal non significatif de PSI-BLAST. Pour qu'elle puisse être détectée, il faudrait que notre méthode permette d'inclure de façon itérative des homologues lointains détectés lors d'un premier filtrage.

Au sein des 68 autres séquences du signal non significatif filtrées par notre approche et ne possédant pas une architecture de type Rad9, nous trouvons 19 séquences qui sont probablement de vrais homologues lointains de la portion de Rad9 étudiée.

- Quinze contiennent un domaine Tudor ou PWWP (ces deux domaines partagent un même repliement), qui s'aligne avec la région 779-831 de Rad9, soit la région du premier tonneau β observé par RMN. Les e-value correspondant aux alignements PSI-BLAST initiaux de Rad9 avec ces séquences sont réparties sur l'ensemble de la plage étudiée : elles vont de 6 à 650. Certains des homologues lointains ainsi identifiés correspondent à des fragments de protéines de la réparation des dommages de l'ADN.

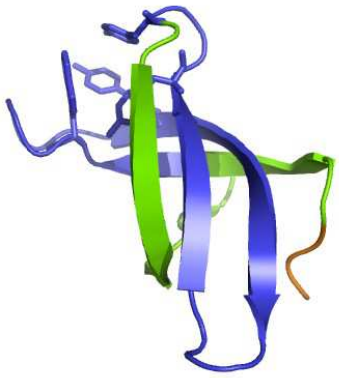
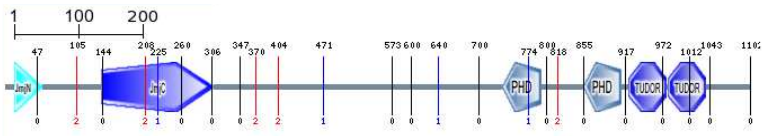


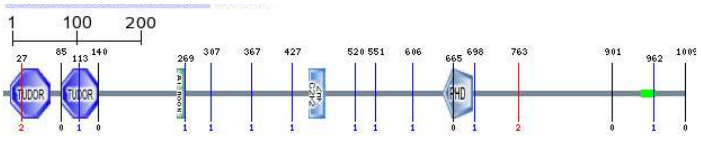
Figure 66 : Structure 3D du domaine Tudor de la protéine humaine SMN (Code PDB : 1G5V,). Sont représentés le squelette polypeptidique sous forme de ruban et l'amas de résidus aromatiques caractéristique des domaines Tudor et PWWP . En orange figure la région N-terminale jamais retrouvée dans nos recherches, en vert la région alignée avec Rad9 dans le cas de plusieurs séquences homologues, et en bleu la région systématiquement incluse dans l'alignement avec Rad9

- Trois sont des enzymes possédant l'architecture suivante :



Le deuxième domaine Tudor de ces enzymes s'aligne avec la région 793-831 de Rad9, c'est-à-dire le premier tonneau β. Ce résultat est là encore cohérent avec l'analyse RMN de la portion de Rad9 étudiée. Les e-value correspondant aux 3 alignements PSI-BLAST étaient initialement de 435, 477 et 485. Les trois fragments identifiés correspondent à des protéines impliquées dans la modification des histones.

- Une séquence possède un tandem de Tudor associé à un doigt de zinc et un domaine PHD. Son architecture est la suivante :



Les deux domaines Tudor sont dans ce cas contenus dans un domaine Ndr de PFAM. Le deuxième domaine Tudor s'aligne avec la région 793-818 de Rad9, c'est-à-dire le premier tonneau β. La e-value initiale correspondant à cet alignement était de 505.

Ainsi, notre approche a permis de prédire la présence d'un domaine Tudor au sein de la région 779 à 831 de Rad9. La même prédiction est réalisée par Superfamily (figure 64). En revanche, le deuxième domaine observé par RMN n'est pas reconnu. Il possède une insertion d'une vingtaine de résidus au niveau d'une boucle qui forme une décoration sur le domaine Tudor et gène probablement sa reconnaissance. Il me faut maintenant réitérer la procédure sur un alignement enrichi en homologues lointains pour tenter d'aller plus loin dans la détection des homologues de cette portion de Rad9.

Enfin, mis à part les domaines Tudor déjà identifiés, il reste 49 séquences à analyser dans le signal non significatif (49/70=70% du bruit), ce qui est cohérent avec le pourcentage de faux positifs générés par notre approche. Certaines de ces séquences correspondent clairement à des faux positifs. En particulier, 11 d'entre elles possèdent un profil obtenu à partir de 2 séquences uniquement. Parmi les 14 séquences de levure, 6 ont un profil restreint à 2 séquences, et 6 autres s'alignent avec Rad9 au niveau de fragments que SMART ou parfois seulement HHSearch prédisent comme des domaines WD40 . La Figure 67 montre la région du « propeller » des domaines WD40 qui s'aligne avec Rad9 (en violet). La structure 3D de cette région contient uniquement des brins β, au même titre que les domaines Tudor ou PWWP. Cette similitude structurale peut être à l'origine du bon score de l'alignement de Rad9 avec des domaines WD40. Le grand nombre de WD40 présent chez les levures (11000 WD40 contre 15 TUDOR et 10 PWWP; cf SMART) nous semble une explication quant au nombre élevé de faux-positifs obtenus.

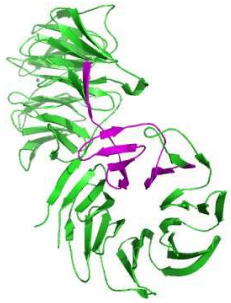


Figure 67 : Structure 3D de la protéine de C.elegans *homologue à la protéine de levure AIP1 interagissant avec l'actine (Code PDB : 1NR0,). En violet figure la région alignée avec Rad9.*

V.3.2.Xrs2, identification de domaines BRCT très divergents.

Xrs2 est une protéine essentielle à la réparation des cassures double-brin de l'ADN chez S.cerevisiæ. Elle fait partie du complexe composé des protéines Mre11-Rad50-Xrs2 (nommé MRX) qui joue un rôle fondamental dans la reconnaissance initiale des cassures et dans l'amplification du signal déclenchant le processus de réparation (Figure 68). Chez l'homme, il existe un homologue Nbs1 fortement divergent en séquence (18 % identité) qui s'associe également en complexe avec les homologues humains Rad50 et Mre11 .

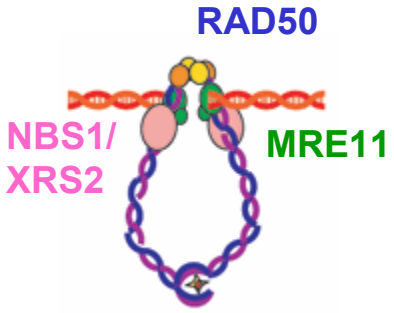


Figure 68 : Schéma indiquant le positionnement approximatif des trois partenaires du complexe MRX (chez S. cerevisiae) et MRN (chez l'homme) lié à une cassure double dans l'ADN. Le complexe fonctionnerait en dimère articulé au niveau d'un domaine crochet de Rad50 (étoile rouge). Xrs2 (ou Nbs1) est en rose, Mre11 en vert, le domaine N-ter de Rad50 en orange et le domaine coiled-coil en bleu/violet.

Au plan de l'organisation en domaines, les deux protéines Xrs2 et Nbs1 sont prédites comme ne comportant qu'un domaine en commun, un domaine FHA dans la région N-terminale (Figure 69). Expérimentalement, des petites régions dans la partie C-terminale des deux protéines ont été montrées comme impliquées dans l'interaction avec Mre11 et ATM .

Chez l'homme, 90 % des patients atteints du syndrome de Nijmegen possèdent une délétion d'un nucléotide en position 657 dans le gène Nbs1 (Figure 69) provoquant des symptômes tels que la microcéphalie, des instabilités chromosomiques, la radiosensibilité, des déficiences immunitaires et une forte prédisposition au cancer . Dans les cellules de mammifères, la délétion de Nbs1 est létale. Dans le cas du syndrome de Nijmegen, la délétion du nucléotide induit une cassure de la protéine Nbs1 au résidu I221 mais ne provoque pas de létalité. Un premier fragment p26 est transcrit et un site alternatif de transcription permet la synthèse du second fragment de la protéine p70 à des niveaux d'expression toutefois inférieurs à ceux de la protéine sauvage.

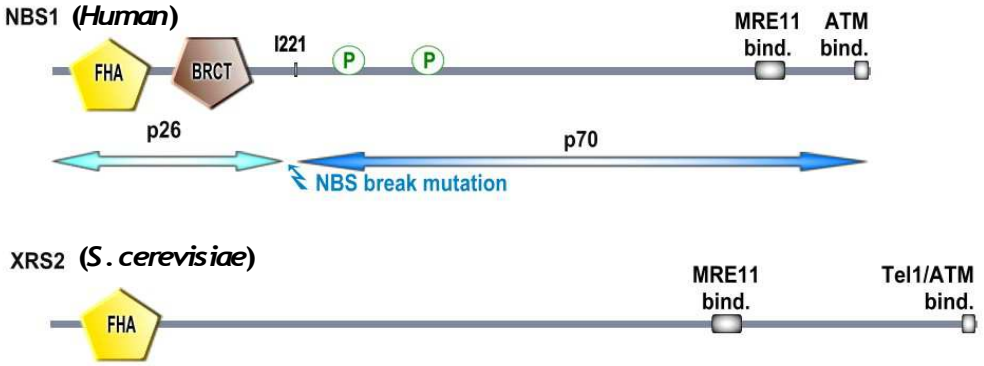


Figure 69 : Représentation des domaines FHA et BRCT identifiés dans les protéines Xrs2 et Nbs1 par les méthodes de détection de domaines. Les carrés gris indiquent les régions expérimentalement montrées comme interagissant avec Mre11 et Tel1/ATM Les ronds vert indiquent l'emplacement des sites de phosphorylations importants dans Nbs1. Une délétion du nucléotide 657 dans l'exon 5 est responsable du syndrome de Nijmegen et induit une cassure au résidu I221.

L'analyse de séquence manuelle que j'ai effectuée sur les protéines Xrs2 et Nbs1 au cours de ma thèse a permis de détecter de façon statistiquement significative un double domaine BRCT. La construction d'un alignement fiable, ainsi que la modélisation à très basse identité (~ 15 % identité) du module double BRCT dans les deux protéines a été effectuée en collaboration avec E. Becker et est détaillée dans la publication en annexe de ce manuscrit . L'analyse des modèles a permis de proposer que les deux domaines BRCT constituent un module de reconnaissance des résidus phosphosérines. Nous avons par ailleurs montré que la mutation du syndrome Nijmegen dans Nbs1 aurait pour effet de couper en deux le double domaine BRCT tout en maintenant l'intégrité de la structure 3D de chacun des domaines isolés. Dans le cadre de ce chapitre, nous ne reviendrons pas sur les détails de l'analyse qui peuvent être consultés dans la publication jointe. Ici, nous avons exploré en détail l'ensemble

des séquences sélectionnées par la stratégie de double filtrage pour évaluer si la détection du double BRCT dans Xrs2 aurait pu être identifiée de façon automatique par notre procédure.

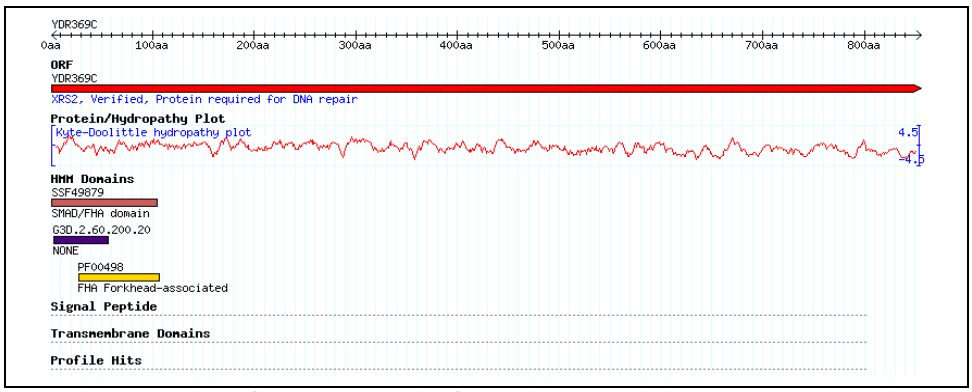


Figure 70 : Représentation schématique (extraite de la base de données SGD) des domaines identifiés dans Xrs2 à partir de différentes méthodes de détection de domaines par des techniques HMM. Les initiales des codes de domaines indiquent la base de donnée source utilisée (SSF : superfamily ; PF : Pfam ; SM :SMART ; PTHR : Panther)

Comme indiqué Figure 70, le domaine BRCT de Xrs2 n'est identifié dans aucune des bases de données d'analyse de domaines actuelles (PFAM, SMART, PANTHER, Superfamily). Notons que le programme Superfamily dans sa version la plus récente utilise le programme PRC de comparaison HMM/HMM . La procédure de sélection par prédiction de structures secondaires suivie de deux comparaisons profil/profil permet de filtrer 77 séquences à partir d'un signal non significatif qui contenait environ 1000 séquences. Sur ces 77 fragments de séquences, 32 (~40 %) contiennent effectivement un domaine BRCT. Seul le premier BRCT de Xrs2 est détecté, le second étant trop divergent pour être aligné dans le profil de PSI-BLAST. La e-value minimale obtenue par PSI-BLAST pour un vrai positif dans lequel la présence d'un BRCT était connue par PFAM est de 15. Les e-value des autres vrais positifs obtenus par PSI-BLAST peuvent atteindre 489, soulignant l'intérêt de récupérer des séquences non significatives jusqu'à des e-values très élevées.

Concernant les faux positifs, deux cas de figure peuvent être distingués :

(i) 19 faux positifs (25 %) associés à des profils constitués de plusieurs séquences relativement divergentes. Nous les considérons comme de réels faux positifs équivalents à ceux détectés dans les chapitre III et IV.

(ii) 26 faux-positifs (34 %) associés à des profils qui ne rassemblent pas plus de deux ou trois séquences (généralement très similaires). Dans ce cas, les outils de prédiction profil/profil sont très peu fiables puisqu'on se ramène à une situation de comparaison profil/séquence. Il s'agit clairement de cas qui pourraient être filtrés en amont de la comparaison profil/profil afin de limiter le nombre de faux

positifs. Ces cas de figures étaient très rares dans les tests effectués aux chapitres III et IV puisque toutes les séquences du signal non significatif étaient associées à une structure de domaine capable *a priori* de s'aligner avec un nombre significatif de séquence permettant la construction d'un profil de qualité.

En conclusion, la stratégie d'exploration du signal non significatif se révèle efficace pour identifier la présence d'un des deux domaines BRCT (Figure 71). Le ratio entre vrai positifs et faux positifs au sein du signal sélectionné apparaît néanmoins sensiblement différent de ce qui avait été calculé au chapitre IV (les vrais positifs constituaient en moyenne près de 80 % du signal filtré). Cependant, si on ne tient pas compte des cas de profils appauvris mentionnés dans le point (ii) ci-dessus, la méthode permet de rassembler dans le filtrat près de 70 % de vrai-positifs en accord avec les résultats obtenus sur la base d'apprentissage. Au plan biologique la détection du second domaine BRCT a permis de proposer une nouvelle hypothèse pour interpréter l'effet délétère de la mutation sur la fonction de la protéine Nbs1 (cf Article en annexe).

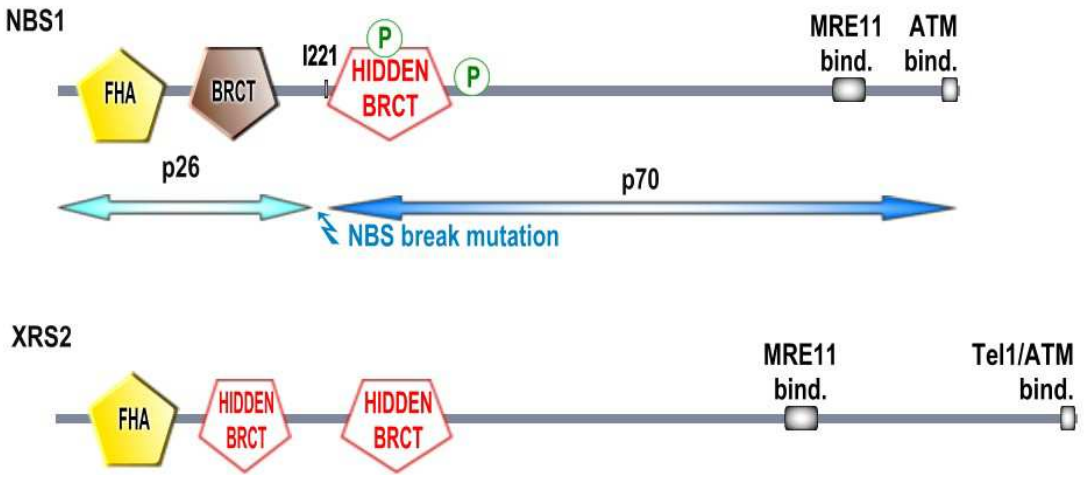


Figure 71 : Emplacement des domaines BRCT prédits dans les séquences de Nbs1 et de Xrs2 .

V.3.3.Nej1, recherche de l'homologue humain.

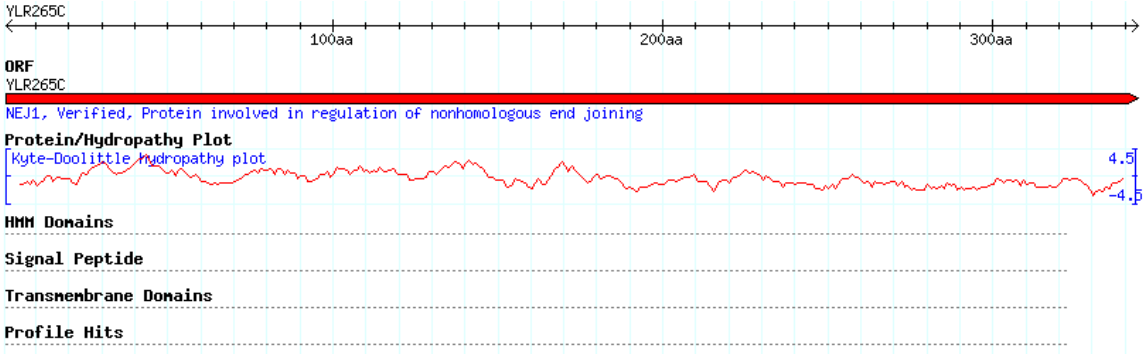


Figure 72 : Représentation schématique des domaines identifiés dans Set1 à partir de différentes méthodes de détection de domaines par des techniques HMM. Les initiales des codes de domaines indiquent la base de donnée source utilisée (SSF : superfamily ; PF : Pfam ; SM :SMART ; PTHR : Panther)

Nej1 (également nommée Lif2) est un cas d'étude particulièrement intéressant et stimulant pour illustrer les limites et les enjeux que peuvent constituer la détection d'homologies lointaines. A priori Nej1 n'est composée d'aucun domaine connu (Figure 72). Suite à l'identification de son rôle dans le NHEJ (Non-Homologous End Joining) par Stéphane Marcand au SBGM/CEA Saclay , plusieurs études en collaboration avec Isabelle Callebaut avait tenté en 2002 d'identifier un homologue de Nej1 chez les eucaryotes supérieurs. Malgré des recherches poussées utilisant les différents outils d'analyse de séquence alors disponibles, aucun indice n'avait permis de récupérer des homologues dans des espèces plus éloignées de *S. cerevisiae* que *K. Lactis*. Cette observation laissait supposer que le système NHEJ avait hérité d'un mécanisme de régulation particulier dans les levures apparentées à *S. cerevisiae* qui n'avait pas été conservé au cours de l'évolution. Récemment, deux équipes ont simultanément identifié une protéine humaine, nommée Cernunnos, jouant un rôle dans le NHEJ . En repartant de la séquence de cette nouvelle protéine humaine et en recherchant des homologues lointains dans le signal non significatif du logiciel PSI-BLAST, I. Callebaut est parvenue à remonter à la protéine Nej1 de levure. De plus, son analyse a permis de démontrer que les protéines de la famille Nej1 sont probablement des homologues lointains de la famille des protéines XRCC4 . Ce cas illustre la difficulté de détection des homologues lointains dans certains cas de divergence très rapide.

Nous avons cherché à caractériser si la stratégie de filtrage du signal non significatif aurait permis de résoudre cette question et si de façon plus générale les méthodes de prédiction profil/profil aurait pu aider à résoudre l'énigme de l'homologue humain de Nej1 en partant de la protéine de levure. La construction du profil de référence à partir de la protéine Nej1 permet après 4 itérations de rapatrier plusieurs séquences homologues dans des levures apparentées à *S. cerevisiae* (Q06148) telles que *K. lactis* (Q6CJ96) et *A. gossypii* (Q74ZD6). Les identités de séquences avec les homologues de ces espèces sont faibles de l'ordre de 20 %. Pour les espèces plus éloignées de *S. cerevisiae* telles que *D. hansenii* (Q6BKD1) aucune séquence n'est retrouvée dans le signal significatif du logiciel PSI-BLAST.

Lorsque l'on analyse maintenant les 34 portions de séquences proposées comme homologues par notre procédure, il apparaît que 24 d'entre elles (71%) ont des profils basés sur 2 ou 3 séquences. Parmi les 10 séquences restantes, on ne trouve que des séquences de parasites, plantes ou mammifères. Ces séquences sont alignées sur le fragment 150-260 de Nej1, qui est prédit en hélice α, et qui correspondrait aux longues hélices super enroulées prédites pour la structure de Cernunnos .

En revanche, dans le signal non significatif non remonté par notre procédure, une séquence de *D. hansenii* (Q6BKD1) est retrouvée avec une e-value de 6.82. Il s'agit de l'homologue lointain de Nej1 identifié en 2006 grâce à la découverte de la protéine humaine Cernunnos. Cette séquence possède un Qsecpred de 32 % avec la séquence de *S. cerevisiae* supérieur au seuil de 20 % défini au chapitre III, et a donc été présélectionnée pour être testée par les deux méthodes de comparaisons profil/profil. COMPASS aligne 54 positions avec une e-value de 4.10^{-2}, HHsearch aligne 76 positions avec une e-value de 1.2. Le seuil de COMPASS ayant été fixé à 10^{-3}, ces valeurs sont trop élevées pour que la séquence de *D. hansenii* soit sélectionnée dans l'ensemble des séquences d'homologues lointains putatifs. Pourtant, les valeurs de e-value obtenues par les deux programmes sont relativement faibles. Ces résultats sont informatifs sur les limites de notre stratégie à son stade de développement actuel. L'utilisation d'un seuil fixe pour déterminer les limites de détection de COMPASS et de HHsearch n'est probablement pas la meilleure approche pour limiter le nombre de faux négatifs. Dans un cas comme celui-ci où la e-value de HHsearch est largement inférieure au seuil de 75, on pourrait imaginer que le seuil associé pour la sélection par HHsearch soit moins strict. Nous avons vu au chapitre IV (cf page 120) que le croisement entre les deux méthodes permettait de réduire drastiquement le nombre de faux-positifs mais induisait une perte de séquences d'homologues lointains de l'ordre de 10 % (une fois ôtés les homologues lointains qui possédaient un Qmod nul). Dans le cas testé ici, ces 10 % contiennent précisément la séquence qui nous aurait permis de remonter jusqu'à l'homologue humain de Nej1. Ainsi apparaît une limitation de l'utilisation de deux seuils indépendants pour les méthodes COMPASS et HHsearch.

V.4. Apport de la méthode de filtrage aux quelques exemples étudiés au laboratoire.

L'analyse préliminaire des résultats obtenus par notre procédure sur 100 séquences de protéines impliquées dans la signalisation et la réparation des dommages de l'ADN, ainsi que l'analyse détaillée de trois exemples issus du laboratoire, mettent en évidence plusieurs défauts de la procédure qui n'avaient pas été révélés par le test effectué sur les protéines de la SCOP.

Tout d'abord, dans 16% des cas pour Rad9, 34% des cas pour Xrs2, et 71% des cas pour Nej1, les homologues putatifs selectionnés par notre procédure ont un profil basé sur moins de 3 séquences. La stratégie mise au point revient alors à comparer le profil de la séquence d'intérêt avec des séquences et non plus des profils, ce qui n'apporte pas *a priori* plus d'informations que la seule utilisation de PSI-BLAST. Les e-values calculées par les programmes de comparaison profil/profil ne semblent pas prendre en compte la pauvreté ou la richesse d'un profil. Notre expérience montre néanmoins que c'est une source majeure de faux positifs. Ces méthodes ayant été calibrées sur des bases de données telles que la SCOP, elles n'ont probablement que rarement rencontré des cas où les profils calculés étaient pauvres en séquences. En première approximation, il apparaît donc que les profils pauvres en séquences devraient être écartés avant l'étape de prédiction profil/profil.

Nous avons observé qu'il était possible de repérer rapidement parmi les séquences sélectionnées par notre procédure, celles qui possédaient la même architecture en domaines et/ou la même fonction que la séquence d'intérêt. Dû à leur faible similarité avec la séquence d'intérêt, ces homologues lointains n'ont pas été inclus dans le profil initial. Pour que la procédure soit plus sensible, il faudrait que notre méthode puisse inclure de façon itérative ces homologues lointains dans le profil initial associé à la séquence d'intérêt.

Enfin, ces exemples ont permis de montrer que de vrais homologues lointains sont perdus lors de notre filtrage alors que le score de la première méthode profil-profil est largement significatif (par rapport aux critères définis au chapitre IV) et le score de l'autre méthode est juste sous le seuil. Dans ces cas-là, on pourrait imaginer un système de classification basé sur un apprentissage supervisé à partir d'une base de données référence (réseau de neurone, SVM) afin d'affiner notre capacité de détection.

Chapitre VI : Discussion générale, Conclusions et Perspectives

VI.1. Contexte scientifique ayant stimulé le développement de notre approche

L'objectif de mon doctorat était de développer une méthode d'analyse des séquences protéiques permettant de cribler, le plus efficacement possible, les alignements non significatifs produits par le logiciel PSI-BLAST et d'identifier ainsi des homologies lointaines entre séquences. La description des protéines sous forme de modules fonctionnels a au cours des dix dernières années a profondément modifié la stratégie d'annotation et d'analyse structurale et fonctionnelle des protéines. Une des difficultés rencontrées par les expérimentalistes repose sur la détection de ces modules et sur leur délimitation. L'outil que j'ai mis au point intéresse en premier lieu la biologie structurale puisqu'il permet d'aider à identifier et à délimiter des modules structurés au sein des protéines de grande taille. Son champ d'application s'étend également aux autres domaines de la génétique ou de la biologie cellulaire car l'identification de modules fonctionnels est souvent un pré-requis à la dissection de la fonction d'une protéine et à la recherche de partenaires.

Deux évolutions majeures en biologie ont suscité le développement et l'automatisation de telles approches. Tout d'abord, l'explosion du nombre de génomes séquencés, particulièrement dans le règne des *fungi*, augmente énormément les possibilités de découvrir des relations, jusqu'alors cachées, entre les séquences de levures et celles d'eucaryotes supérieurs. Les trois exemples mentionnés au chapitre VI fournissent une illustration des enjeux de ces détections. De plus, au plan méthodologique, on peut noter que chaque décennie a été accompagnée par une évolution majeure des outils d'analyse. Dans les années 80, les méthodes heuristiques telles que BLAST et FASTA ont accélérées les recherches d'homologues à grande échelle sur les bases de données. Les années 90 ont vu l'essor des approches de comparaison profil/séquence ou HMM/séquence, telles que HMMER et PSI-BLAST. Sans doute, les années 2000 seront marquées par le développement de méthodes de comparaison profil/profil (ou HMM/HMM) de plus en plus performantes.

A mesure que les limites de détection de l'homologie sont repoussées, la prise en compte de l'information structurale déduite de l'analyse de la séquence devient de plus en plus importante. Au sein d'un alignement de séquences, les corrélations existant entre les acides aminés et reflétant les contraintes de maintien de la structure tridimensionnelle

constituent des marqueurs évolutifs très utiles pour reconnaître les homologies lointaines. Parmi ces informations, la nature des structures secondaires nous a paru particulièrement utile à prendre en compte. Elle peut être prédite rapidement et avec une précision relativement informative. De plus, cette prédiction possède un grand champ d'applications puisqu'au contraire des techniques de threading elle peut être appliquée à des régions de protéines pour lesquelles aucune structure homologue n'existe.

VI.2.Rappel des principaux résultats

Lors de ma thèse, j'ai mis au point une stratégie permettaient de retrouver des homologues lointains dans le signal non-significatif de PSI-BLAST. Afin d'évaluer cette stratégie, j'ai tout d'abord vérifié la présence de séquences homologues parmi ce signal (cf chapitre II). J'ai observé que les alignements non significatifs donné par le logiciel PSI-BLAST dans l'intervalle (10^{-3}, 1000) rassemblent un ensemble de séquences enrichie en homologues lointains. J'ai montré que des alignements présentant des e-values fortement non significatives (entre 100 et 1000) sont associés à une réalité structurale dans environ un tiers des cas. Enfin, j'ai constaté que parmi le signal non significatif du logiciel PSI-BLAST, les alignements de grande taille sont généralement de mauvaise qualité : la taille moyenne des alignements corrects à plus de 90% est de seulement trente résidus.

A partir de cette première analyse, j'ai choisi de rechercher des homologues lointains dans le bruit de fond de PSI-BLAST défini dans un intervalle de e-values compris entre 0,001 à 1000. J'ai alors isolé un jeu de données test de 200 séquences issues de la banque SCOP10, pour lesquelles plus de 10 séquences de la même superfamille sont retrouvées dans le signal non significatif de PSI-BLAST après une recherche sur la banque SCOP40. C'est sur ce jeu test que j'ai effectué la mise au point de ma procédure de recherche d'homologues lointains.

J'ai ensuite montré que les prédictions de structure secondaire permettaient de filtrer efficacement le signal non significatif de PSI-BLAST (cf chapitre III). La comparaison d'un point de vue local et global des prédictions de structures secondaires entre la séquence d'intérêt et les séquences potentiellement homologues a montré que ces deux approches étaient en mesure de filtrer le signal non significatif. Toutefois, l'approche globale présente des résultats uniquement pour les protéines prédites à plus de 70% en hélices α ou en

feuillets β. De plus, elle suppose que la séquence étudiée ne comporte qu'un domaine. Pour la suite de cette étude, nous avons donc préféré l'utilisation d'une approche locale des prédictions de structures secondaires. Cette approche a permis un filtrage rapide d'environ 40% des séquences non homologues du signal non significatif de notre jeu test. De plus, la majorité des séquences homologues perdues lors du filtrage sont associés à un alignement non-significatif fortuit ne présentant aucune réalité structurale.

Pour aller plus loin dans le filtrage des séquences du signal non-significatif de PSI-BLAST, j'ai utilisé des méthodes de comparaison profil/profil COMPASS et HHsearch (cf chapitre IV). Un des facteurs limitant dans l'utilisation des comparaisons profil/profil à grande échelle réside dans le temps de calcul nécessaire à la construction des profils associé à chaque protéine du signal non significatif. L'originalité de l'approche développée ici consistant à croiser les résultats de deux méthodes de comparaisons profil/profil permet pour le coût de la construction d'un seul profil de profiter de la précision de deux algorithmes différents de comparaison profil/profil. J'ai montré que ces deux méthodes étaient complémentaires et qu'un croisement de leurs résultats permettait d'augmenter la spécificité des résultats obtenus indépendamment. Les deux méthodes présentent en effet très peu de recouvrements entre leurs faux-positifs et offre une sensibilité tout à fait intéressante pour identifier un nombre conséquent de séquences d'intérêt. Cette intégration de différents outils d'analyse de séquences permet de traiter dans un temps moyen de 2h (cluster de 10 Processeurs Intel Xeon 3GHz) l'ensemble des séquences présentes dans le signal non significatif d'une séquence test. Ainsi, environ 90% des séquences non homologues présentes parmi le signal non significatif de notre jeu test ont été éliminées car leurs scores COMPASS et/ou HHsearch étaient supérieurs au seuil fixé. Là encore, j'ai vérifié que l'approche choisie ne conduisait qu'à l'élimination d'un faible nombre d'homologues lointains ne correspondant pas à un alignement fortuit (environ 20 %).

Nous avons discuté au chapitre V que la validation de la procédure sur une base de données de séquences dont les structures étaient déjà connues n'était pas forcément représentative de tous les cas de figures rencontrés lors d'une recherche d'homologues lointains sur une base de donnée telle que la nr. Dans ce contexte, l'analyse de cas réels extraits d'un ensemble de protéines de la réparation et de la signalisation des dommages de l'ADN nous a permis de tester les potentialités réelles de l'approche. Aujourd'hui, l'analyse

préliminaire des résultats obtenus par notre procédure sur 100 séquences de protéines impliquées dans la signalisation et la réparation des dommages de l'ADN, ainsi que l'analyse détaillée de trois exemples issus du laboratoire, ont permis de valider la stratégie utilisée :

- dans 36% des cas, des informations potentiellement nouvelles sur la composition en domaines des portions étudiées ont été trouvées.
- dans 27% des cas, des domaines cachés (non identifiés dans la PFAM) mais déjà détectés dans des protéines de fonctions similaires ont été retrouvés.

Toutefois, ces analyses ont aussi mis en évidence des défauts dans notre procédure, que nous souhaitons aujourd'hui rapidement corriger et m'ont stimulé pour apporter de nouvelles améliorations en particulier, pour améliorer à court et moyen termes la spécificité de la procédure.

VI.3.Comparaison avec les résultats de la littérature.

A l'époque où mon travail de thèse a été initié en décembre 2003, le domaine de la recherche d'homologues lointains par des approches automatiques était dominé par les techniques de « threading ». Comme mentionné plus haut, l'inconvénient du « threading » est qu'il suppose l'existence d'une structure tridimensionnelle correspondant à la portion de séquence analysée. Les premières méthodes de comparaison profil/profil ont été publiées à cette période et j'ai, au début de ma thèse, évalué l'intérêt des différentes approches telles que COMPASS , COACH ou HHsearch . COMPASS et HHsearch ont été sélectionnées pour leur rapidité et parce qu'elles intégraient en plus d'une procédure d'alignement des profils, un paramètre statistique (e-value ou probabilité) qui permettait d'évaluer la vraisemblance des alignements. Nous avons vu au chapitre IV qu'il existait néanmoins une disparité entre les paramètres statistiques calculés par les deux approches. Au cours de la discussion, je reviendrai brièvement sur cette particularité. Dans la suite, je mentionnerai également différents travaux publiés en 2006 qui s'inscrivent dans une démarche similaire à celle développée au cours de mon doctorat. Je chercherai à dégager les points originaux de ces travaux, afin de décrire ce qui les distinguent de la démarche entreprise dans mon étude.

VI.3.1.Comparaison avec les résultats publiés sur les logiciels HHsearch et COMPASS

Deux points importants ont été notés lors de l'analyse comparée des capacités discriminantes des deux programmes COMPASS et HHsearch :

- Les e-value obtenues avec COMPASS favorise l'existence de faux-positifs. Pour des e-value inférieures à 10^{-3} on observe ainsi encore un nombre important de faux-positifs. En revanche, la e-value du programme HHsearch apparaît très stricte.
- Les différences observées entre les résultats du programme COMPASS et HHsearch sont étonnantes en regard des données publiées sur la sensibilité des deux méthodes. (courbe de sensibilité des logiciels PSI-BLAST, COMPASS, HHnoss et HHss tracée publié par J. Soeding et présentée en introduction Figure 19 page 53).

I.1.1.23Comparaison des calculs de e-value

Les modes de calcul des e-values entre les deux méthodes sont radicalement différents. Dans le cas de COMPASS, comme indiqué dans l'introduction, le mode de calcul de la e-value est très similaire à celui utilisé par PSI-BLAST. Ce calcul a été optimisé pour favoriser la sensibilité de la méthode . La e-value est calculée à partir d'une formule analytique qui prend en compte la longueur des séquences alignées et les paramètres λ et K. Ces derniers paramètres caractérisent la distribution des scores attendus lors de l'alignement d'une séquence avec l'ensemble des séquences d'une base de données. Les paramètres nécessaires pour calculer λ et K ont donc été obtenu une fois pour toute au moment du développement du programme et ne sont pas modifiés par la suite. En revanche, avec HHsearch, les paramètres λ et K sont définis à chaque étape de calibration d'un profil HMM. Bien que plus coûteuse en temps de calcul, cette méthode semble fournir des e-values en meilleur accord avec le nombre de faux-positifs effectivement obtenus. A ce propos, J. Soeding mentionne sur le site de HHsearch les résultats d'une analyse sur la précision de PSI-BLAST (http://toolkit.tuebingen.mpg.de/hhpred/help_ov). Sur un grand ensemble de séquences, les résultats de PSI-BLAST (utilisé avec les paramètres standards, jusqu'à 8 itérations avec des e-values d'inclusion de 10^{-4}) comptent 100 fois trop de séquences non homologues dans l'ensemble des séquences supposées homologues pour une e-value calculée par PSI-BLAST inférieure à 10^{-4}.

I.1.1.24Hiérarchie dans les performances des programmes

Il reste néanmoins surprenant que les résultats de HHsearch ne soient pas supérieurs à ceux de COMPASS en regard de ce qui a été publié (Figure 19 page 53). Notons que dans la

publication originale, J. Söding mentionnait effectivement une perte de fiabilité de la méthode HHsearch pour des homologues très lointains correspondant au cas de figure exploré ici . Elle n'était néanmoins pas aussi importante que celle obtenue dans nos tests. Une des différences entre notre étude et son ensemble test provient aussi de la longueur des séquences testées pour effectuer la comparaison. Dans son cas, J. Söding utilise l'intégrité de la longueur des deux séquences extraites de SCOP pour construire les profils et les comparer ensuite entre eux avec HHsearch. En revanche, nous restreignons la taille de la séquence prise en compte pour le second profil à la longueur de l'alignement identifié dans le signal non significatif de PSI-BLAST. Les profils résultants sont donc probablement de moins bonne qualité que ceux testés par J. Söding. Ce point souligne l'intérêt de ne pas se restreindre aux résultats des benchmarks pour décider de l'utilisation d'un logiciel ou d'un autre mais de tester chaque algorithme dans les conditions d'utilisation (telles qu'ici pour des applications sur des fragments de domaines).

VI.3.2.Le serveur HHsenser.

Récemment, le groupe d'Andrei N.Lupas a publié la description du serveur HHsenser (http://toolkit.tuebingen.mpg.de/hhsenser), qui permet d'enrichir des profils PSI-BLAST en vrais homologues tout en étant très stricts sur le filtrage des faux positifs . HHsenser part d'une séquence unique ou d'un alignement multiple et calcule deux alignements multiples, l'un très stringent et l'autre plus permissif. Pour cela, il effectue un PSI-BLAST sur la séquence d'intérêt avec un seuil de 10^{-3} , trouve une première génération d'homologues potentiels, puis reprend chacun de ces homologues (ou tout au moins un échantillon relativement divergent de ces homologues) et relance PSI-BLAST afin d'obtenir des alignements de deuxième génération. Ces alignements sont comparés avec l'alignement stringent de la séquence de départ en utilisant HHsearch. Des e-values et des P-values sont alors calculées. Si ces valeurs se trouvent sous un certain seuil, l'alignement est ajouté à l'alignement permissif de la séquence de départ. Et si ces valeurs sont sous un deuxième seuil, alors l'alignement est ajouté à l'alignement strict de la séquence de départ.

HHsenser utilise une valeur de 1 pour la e-value seuil ce qui limite le nombre de séquences analysées. Rappelons que dans notre cas l'utilisation couplée de deux méthodes de comparaison profil/profil nous permet d'envisager le traitement des e-value jusqu'à 1000. Cependant, HHsenser tire profit de la transitivité de la relation d'homologie, ce qui lui permet d'explorer avec HHsearch un nombre intéressant de couple de séquences. Toute la question

est donc de savoir si pour toute séquence d'une base de données, il existe une séquence de « pontage » qui permet de passer d'un homologue lointain à l'autre sans jamais avoir à dépasser une e-value PSI-BLAST de 1. Pour la suite, il sera tout à fait intéressant de comparer si notre stratégie permet effectivement une plus grande efficacité.

VI.3.3.La méthode FOLDpro

Le serveur FOLDpro (http://www.igb.uci.edu/servers/psss.html) cherche à déterminer le repliement le plus probable pour une séquence donnée . Son objectif est donc différent du nôtre. Cependant, il utilise une combinaison de méthodes dans son calcul de score, et en cela il se rapproche de notre procédure. FOLDpro exploitent les scores obtenus par un grand nombre de méthodes de comparaison séquence/séquence, séquence/profil, profil/profil (dont COMPASS et HHsearch), et des informations sur la composition des séquences et des structures 3D, afin de calculer un score global correspondant à l'adéquation entre une séquence et un repliement. Ce score global résulte d'une fonction mathématique dont les paramètres ont été optimisés par un processus d'apprentissage de type SVM (Support Vector Machine).

L'approche proposée dans ce travail est intéressante puisqu'elle rejoint nos préoccupations d'améliorer le filtre de sélection des homologues lointains en combinant les scores HHsearch et COMPASS de façon plus optimale que cela n'a été fait au chapitre IV. Là encore une comparaison des résultats obtenus par FOLDpro avec certains cas d'étude que nous avons traités au chapitre V pourra être très intéressante. Néanmoins, il faut noter qu'au dernier CAFASP5, la méthode FOLDpro arrive loin derrière le programme HHpred (correspondant à HHsearch utilisé sur une base de profils extraits de la PDB). Il n'est pas impossible que FOLDpro utilise trop de sources d'informations et que le bruit apporté par certaines méthodes limite les performances du programme.

VI.4.Perspectives d'améliorations de la méthode

VI.4.1.Gains en temps de calcul

L'originalité de l'approche développée ici est tout d'abord d'utiliser les structures secondaires comme un filtre pour éliminer un grand nombre de séquences polluant le signal potentiellement intéressant. Nous avons utilisé un paramètre simple, le Qsecpred, qui permet un traitement efficace des séquences. D'autres stratégies plus sophistiquées, basées sur analyse globale des prédictions de structures secondaires, ont été explorées mais n'ont pas permis d'aboutir à des résultats satisfaisants. L'augmentation de la fiabilité des prédictions de structures secondaires pourrait permettre d'éliminer un plus grand nombre de séquences au départ et donc d'augmenter la rapidité de la procédure. Une des voies d'amélioration de ce filtrage consiste à augmenter la fiabilité des prédictions des structures secondaires à partir de séquences uniques. Ainsi, une méthode basée sur de nouveaux modèles de corrélations entre résidus et des algorithmes récents d'apprentissage, ainsi que sur la représentation HSSM (hidden semi-Markov model) des alignements a été récemment publiée, et pourrait être intégrée à notre procédure . Les nouveaux développements de l'approche HCA sont également tout à fait prometteurs car ils permettent de gagner significativement en fiabilité pour la prédiction en structures secondaires d'un petit nombre d'amas hydrophobes (I. Callebaut, JC. Gely, communication personnelle).

Les résultats obtenus suggèrent plusieurs voies d'améliorations. Pour gagner en temps et en spécificité, il serait tout à fait intéressant de travailler sur des bases de données pré-filtrées. L'utilisation de bases de données telles que la nr dans laquelle aucune séquence ne possède plus de 70 % d'identité avec les autres séquences permettrait de réduire considérablement la taille des ensembles de séquences. Cette stratégie est déjà adoptée sur le site du programme HHpred basé sur l'utilisation de la méthode HHsearch et constitue une voie d'amélioration facile à implémenter (http://toolkit.tuebingen.mpg.de/hhpred).

VI.4.2.Gains en sensibilité et spécificité

L'automatisation de l'approche de criblage des séquences développée dans cette thèse permet d'explorer de nombreuses combinaisons d'analyses. Par exemple, chaque séquence rapatriée au sein du signal significatif pourrait être elle-même utilisée systématiquement comme sonde pour un nouveau criblage du signal non significatif. De plus, la méthode pourrait permettre d'intégrer des homologues lointains de façon stricte en évitant un des

écueils principaux de PSI-BLAST qui est de diverger dès qu'une séquence non homologue est insérée dans un profil.

L'étude menée au chapitre VI nous a également permis d'identifier quelques limites de l'approche qui pourront être facilement surmontées dans la suite du développement du programme. Une première ligne d'amélioration concerne la définition d'un indice de qualité des profils générés, qui devrait permettre de filtrer tous les profils ne contenant pas assez de divergence pour pouvoir amener une prédiction fiable. De plus, en travaillant sur des régions de séquences vierges de toute annotation, nous avons constaté que fréquemment les itérations de PSI-BLAST ne permettaient pas de récupérer un nombre suffisant de séquences dans le signal significatif. L'intégration d'une procédure itérative permettant d'enrichir les profils initiaux en homologues lointains permettra d'augmenter sensiblement le potentiel de l'approche. Enfin, la technique d'évaluation des seuils utilisés pour coupler les prédictions de COMPASS et HHsearch reste assez naïve. Nous avons vu dans le chapitre V que les limites de ce système nous empêchaient d'identifier de façon automatique l'homologue lointain de Nej1 et de remonter ainsi à l'homologue humain. Nous envisageons de raffiner les capacités de détection de la méthode en utilisant des techniques d'apprentissage (telles que la méthode SVM (Support Vector Machine) ou réseau de neurones) pour améliorer la discrimination entre les homologues lointains et les faux positifs générés par PSI-BLAST. Des informations plus générales associées à la phylogénie ou aux fonctions cellulaires de la protéine et de ses homologues potentiels pourront aussi être intégrées. Nous espérons par cette approche découvrir de nouveaux liens entre les protéines de la signalisation des dommages de l'ADN.

VI.4.3.Gains en facilité d'utilisation

La sortie des résultats est pour le moment présentée sous forme de fichiers plats et ne laisse pas beaucoup de place à l'interactivité. Pourtant, L'approche peut gagner énormément si elle permet un post-traitement des données de sortie convivial sous un format de type HTML. L'intuition et la culture biologique de l'expert peut sur un ensemble réduit d'une cinquantaine de séquences s'avérer très complémentaire des approches présentées ici. A titre d'exemple, nous pourrions facilement envisager de coupler l'affichage des résultats filtrés avec des liens permettant de choisir des analyses complémentaires de type HHpred, pour aller plus loin dans l'analyse des séquences les plus intéressantes. Un système de seuils ajustables pour adapter suivant les cas de figure les seuils de filtrage des homologues lointains serait également possible à intégrer assez rapidement.

Pour conclure, les résultats présentés dans cette thèse ouvrent un champ d'investigation important pour la recherche semi-automatisée d'homologies lointaines entre les séquences. En s'appuyant sur les outils parmi les plus performants du moment, l'approche devrait permettre d'accélérer le processus de découverte de nouveaux domaines ou de nouvelles fonctions et stimuler des travaux en biologie structurale et en biologie cellulaire de façon plus générale. Une bonne illustration de ce processus concerne les développements stimulés par la découverte du tandem BRCT dans des protéines très étudiées telles que Xrs2 et Nbs1. Les cristallographes du laboratoire sont actuellement en train de purifier le module FHA-tandem BRCT des deux protéines, notre collaboratrice généticienne M.C. Marsolier-Kergoat (DBJC, CEA Saclay) effectue des recherches de partenaires du module de Xrs2 dans la levure et nous avons initié une collaboration avec N. Lowndes (Galway, Ireland) pour rechercher des partenaires de ce module dans Nbs1 sur un modèle de cellules animales DT40 très prometteur. J'espère que d'autres exemples comme celui-ci pourront être développés à l'avenir grâce à l'aide apportée par l'outil que j'ai développé.

Chapitre VII :BIBLIOGRAPHIE

Chapitre VIII :ANNEXES

VIII.1.SORTIES DU PROGRAMME DE DOUBLE FILTRAGE DES HOMOLOGUES LOINTAINS POUR RAD9, XRS2 ET NEJ1

La première partie de l'annexe présente les sorties du programme d'analyse des homologues lointains annotée pour identifier trois catégories de séquences :

> : désigne les fragments contenant effectivement un domaine homologue lointain de la séquence requête

> : désigne les fragments correspondant a priori à des faux positifs

> : désigne les fragments associés à des profils pauvres. Pas plus de deux ou trois séquences pratiquement identiques

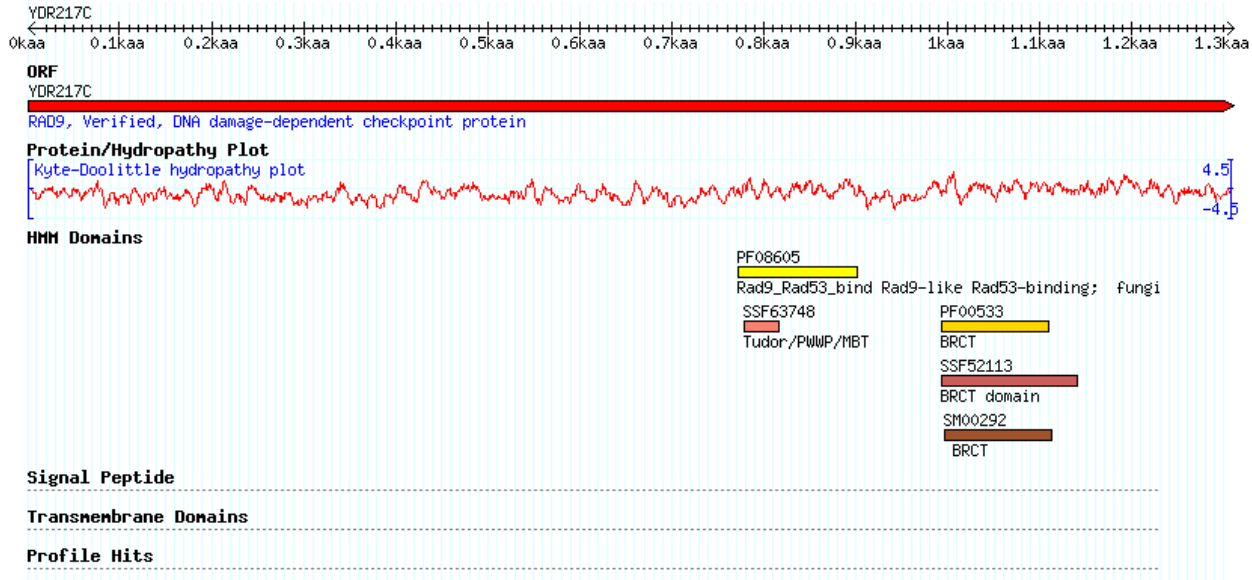
Code de lecture de l'information concentrée dans ces résultats :

```
##############################
P06701@2.info     Code de la protéine d'intérêt@ index du domaine
Sir3              Nom de la protéine d'intérêt
  "Silencing protein that interacts with Sir2p and Sir4p, and histone H3 and H4 tails, to es-
  tablish a transcriptionally silent chromatin state; required for spreading of silenced chro-
  matin; recruited to chromatin through interaction with Rap1p"        YLR442C
                  Annotation SGD
##############################
Lecture des informations :
>NomDomaine | CodeProtDetectée | EvaluePsiBlast | EvalueCOMPASS (seuil evalue=10-
  3) | EvalueHHsearch (seuil evalue=75) | definition | DelimitationsMatchDansSeqRéf |
  NbSeqProfilRéf | NbSeqProfilCible | DelimitationsMatchDansSeqCible
*NomDuDomainePFAM|DélimitationDuDomaine|
  X,SiLeDomaineRecouvreLaRégionSélectionnée

Exemple :
>P06701@2|P70044|3.01318|1.47e-128|1e-10|Cell division control protein 6. [Xenopus
  laevis (African clawed frog)]|(576, 817)|26|467|('192', '396')
*AAA|191-395|X
```

RAD9
70 séquences

Légende des couleurs page 171



>P14737@1|Q6B8Q0|359.166|0.00037|3.4|Hypothetical plastid protein.[Gracilaria tenuistipitata var. liui (Red alga)]|('360','433')
None

>P14737@1|Q7QVU5|554.274|0.00013|5.8|GLP_178_48111_45205.[Giardia lamblia ATCC 50803]|('324','384')
None

>P14737@1|Q66HN6|485.005|0.00018|23.0|Solute carrier family 27 (Fatty acid transporter), member 32.[Rattus norvegicus (Rat)]|(815,842)|7|1155|('454','489')
AMP-binding|80|512|X

>P14737@1|Q4W9Z5|374.466|0.00059|0.65|Hypothetical protein.[Aspergillus fumigatus (Sartorya fumigata)]|('358','381')
None

>P14737@1|Q4CTP8|230.803|0.00024|1.9|Hypothetical protein (Fragment).[Trypanosoma cruzi]|('2471','2528')
None

>P14737@1|Q10451|480.975|8.35e-05|1.2|Hypothetical protein C1093.01 in chromosome I.[Schizosaccharomyces pombe (Fission yeast)]|(774,784)|7|2|('62','72')
PPR|1182-1216|

>P14737@1|Q9NHQ0|221.372|0.00028|2.4|Scribbler.[Drosophila melanogaster (Fruitfly)]|(838,856)|7|24|('916','934')
None

>P14737@1|Q9NGW5|221.372|0.000282|2.4|Brakelessprotein.[Drosophila melanogaster(Fruitfly)]|(838,856)|7|24|('906','924')
None

>P14737@1|Q9V8E6|221.372|9.61e-05|3.1|CG5580-PA, isoformA. [Drosophila melanogaster(Fruitfly)]|(838,856)|7|24|('906','924')
None

>P14737@1|Q4XA04|70.5804|0.000232|0.082|Hypotheticalprotein.[Plasmodium chabaudi]|(772,855)|7|6|('31','110')
Peptidase_C48|105-500|X

>P14737@1|Q4YVS2|10.4441|0.000115|0.1|Hypotheticalprotein(Fragment). [Plasmodium berghei]|(755,855)|7|6|('740','832')
Peptidase_C48|827-1276|X

>P14737@1|Q7XYI5|94.5173|0.00027|10.0|Photosystem IIproteinPsbP. [Chlorarachnionsp.(strainCCMP 621)(Pedinomonas minutissima)]|(795,850)|7|31|('227','281')
PsbP|84-288|X

>P14737@1|O35876|170.919|7.81e-09|0.82|Survivalmotorneuronprotein.[Rattus norvegicus(Rat)]|(784,815)|7|256|('97','125')
SMN|24-289|X

>P14737@1|Q7Q9V5|319.568|0.000139|3.1|ENSANGP00000021762 (Fragment). [Anophelesgambiae str.PEST]|('52','70')
None

>P14737@1|Q8BYP5|449.918|6.11e-05|2.7|(Fragment).[Mus musculus(Mouse)]|('119','137')
None

>P14737@1|Q7SFH0|196.966|0.000422|0.61|Hypotheticalprotein.[Neurospora crassa]|('384','409')
None

>P14737@1|Q9NBW0|221.372|0.000282|2.4|BrakelessA (CG5580-PC, isoformC) (RE69316p).[Drosophilamelanogaster(Fruitfly)]|(838,856)|7|24|('906','924')
None

>P14737@1|P97524|505.666|0.000189|23.0|carrierfamily27 member 2).[Rattus norvegicus(Rat)]|(815,842)|7|1155|('454','489')
AMP-binding|80-512|X

>P14737@1|Q5DCL3|380.768|0.000711|2.4|SJCHGC04417 protein.[Schistosoma japonicum (Bloodfluke)]|('93','131')
None

>P14737@1|Q7QGR9|649.491|0.000164|23.0|ENSANGP00000018261 (Fragment). [Anophelesgambiae str.PEST]|(813,859)|7|1138|('428','478')
AMP-binding|405-481|X

>P14737@1|O04716|306.511|9.14e-06|0.66|DNA mismatch repairproteinMSH6-1 (AtMsh6-1).[Arabidopsisthaliana(Mouse-ear cress)]|(779,821)|7|78|('122','164')
MutS_I|380-496| MutS_III|693-788| MutS_IV|886-977| MutS_V|1028-1273|

>P14737@1|Q4RWG7|505.666|8.04e-05|0.067|(Fragment).[Tetraodon nigroviridis (Green puffer)]|(793,818)|7|167|('184','208')
Ndr|2-276|X zf-C2H2|307-332|

>P14737@1|Q4RWG6|505.666|4.53e-07|0.019|Chromosome undetermined SCAF14988,whole genome shotgun sequence.[Tetraodon nigroviridis(Green puffer)]|(793,818)|7|34|('9','33')
zf-C2H2|132-157|

>P14737@1|Q9NBL3|221.372|0.000282|2.4|Scribblershortisoform.[Drosophila melanogaster(Fruitfly)]|(838,856)|7|24|('906','924')
None

>P14737@1|Q2U1A0|311.668|0.00081|0.71|Predictedprotein.[Aspergillusoryzae]| ('357','380')
None

>P14737@1|Q9VAA9|50.1319|4.82e-05|0.91|CG7946-PA (LD23804p).[Drosophila melanogaster(Fruitfly)]|(797,828)|7|318|('26','58')
PWWP|9-75|X

>P14737@1|Q5VYJ2|476.978|0.000187|2.3|Jumonjidomain containing2C.[Homo sapiens(Human)]|(793,831)|7|23|('949','987')
JmjN|17-62| JmjC|177-293|

>P14737@1|Q7R1S1|261.575|0.000109|1.4|GLP_190_74423_74914.[Giardia lambliaATCC 50803]|('133','153')
None

>P14737@1|Q3UNT5|144.65|0.000133|24.0|(Fragment).[Mus musculus(Mouse)]| ('720','778')
None

>P14737@1|P97801|155.93|1.83e-08|1.1|Survivalmotorneuronprotein.[Mus musculus(Mouse)]|(784,815)|7|255|('96','124')
SMN|23-288|X

>P14737@1|Q7YS86|316.913|3.72e-09|0.7|Survivalmotorneuron (Fragment). [Canisfamiliaris(Dog)]|(784,827)|7|264|('90','132')
SMN|17-283|X

>P14737@1|Q9Y817|109.834|0.000144|7.3|SPBC1105.10protein.
[Schizosaccharomycespombe (Fissionyeast)]|(796,860)|7|2|('341','408')
None

>P14737@1|Q2U9Z5|212.327|2.42e-06|3.0|Predictedprotein.[Aspergillusoryzae]|
(796,867)|7|4|('848','911')
Rad9_Rad53_bind|828-975|X BRCT|1019-1133|

>P14737@1|Q8SX80|221.372|0.000282|2.4|LD21711p.[Drosophilamelanogaster
(Fruitfly)]|(838,856)|7|24|('532','550')
None

>P14737@1|Q9HFK5|33.0318|0.000811|0.68|HypotheticalproteinB11E6.110.
[Neurosporacrassa]|(840,863)|7|9|('354','377')
None

>P14737@1|Q5ATX5|11.5439|4.41e-07|2.0|Hypothetical protein.[Aspergillus
nidulansFGSC A4]|(771,868)|7|4|('1376','1460')
tRNA_int_endo|317-399| Rad9_Rad53_bind|1376-1516|X BRCT|1560-1675|

>P14737@1|Q9ULD9|449.918|8.39e-05|2.7|KIAA1281 protein(Fragment).[Homo
sapiens(Human)]|('416','434')
None

>P14737@1|Q65XF2|407.051|3.32e-05|0.37|HypotheticalproteinOJ1504_G04.1
(PSAG protein).[Oryza sativa(japonicacultivargroup)]|('24','50')
None

>P14737@1|Q4CNU0|353.222|7.44e-05|0.047|Lysosomal/endosomalmembrane
proteinp67,putative.[Trypanosoma cruzi]|(775,808)|7|59|('244','277')
Laminin_A|6-595|X

>P14737@1|Q5JCT1|413.901|3.32e-05|0.37|OSA15 protein.[Oryza sativa(japonica
cultivargroup)]|('24','50')
None

>P14737@1|Q3B8E4|6.65568|5.86e-08|2.2|Hypothetical protein(Fragment).
[Xenopus laevis(Africanclawed frog)]|(784,831)|7|271|('97','143')
SMN|22-287|X

>P14737@1|Q56A10|449.918|6.11e-05|2.7|Zfp608protein.[Mus musculus(Mouse)]|
('393','411')
None

>P14737@1|Q7QGS0|435.149|0.000164|23.0|ENSANGP00000025905
(Ensangp00000018270)(Fragment).[Anophelesgambiae str.PEST]|(813,859)|7|
1138|('564','614')
AMP-binding|541-617|X

>P14737@1|Q4Y197|69.4123|0.00037|0.14|Hypothetical protein (Fragment). [Plasmodium chabaudi]|(772,855)|7|6|('8','87')
Peptidase_C48|82-477|X

>P14737@1|Q6P934|50.9755|1.27e-09|0.23|Survival motor neuron 1. [Brachydanio rerio (Zebrafish) (Danio rerio)]|(783,815)|7|272|('82','114')
SMN|15-279|X

>P14737@1|Q3SYM6|449.918|8.39e-05|2.7|ZNF608 protein (Fragment). [Homo sapiens (Human)]|('394','412')
None

>P14737@1|Q9GRA9|221.372|9.61e-05|3.1|Master of thick veins. [Drosophila melanogaster (Fruitfly)]|(838,856)|7|24|('916','934')
None

>P14737@1|Q5AWZ9|27.2639|0.00082|0.68|Hypothetical protein. [Aspergillus nidulans FGSC A4]|('358','381')
None

>P14737@1|Q61WI5|179.693|0.00090|8.9|Hypothetical protein CBG04400. [Caenorhabditis briggsae]|('20','123')
None

>P14737@1|Q9W6S8|49.7154|3.48e-09|0.46|Survival motor neuron protein 1. [Brachydanio rerio (Zebrafish) (Danio rerio)]|(783,815)|7|271|('84','116')
SMN|15-281|X

>P14737@1|Q872W3|198.617|0.00013|2.6|Hypothetical protein B2G14.050. [Neurospora crassa]|('204','241')
None

>P14737@1|Q3ZTR8|387.175|8.83e-05|16.0|HHL (Fragment). [Homo sapiens (Human)]|(821,850)|7|29|('8','36')
TBC|188-399|

>P14737@1|Q4WKW2|682.834|0.00080|16.0|Hypothetical protein. [Aspergillus fumigatus (Sartorya fumigata)]|('102','114')
None

>P14737@1|Q7QSC0|59.7325|2.26e-05|0.86|GLP_105_27923_22137. [Giardia lamblia ATCC 50803]|('412','481')
None

>P14737@1|Q8CH69|198.617|1.83e-08|1.1|Survival of motor neuron protein (Fragment). [Mus musculus (Mouse)]|(784,815)|7|255|('72','100')
SMN|1-264|X

>P14737@1|Q4I410|633.436|0.00082|0.7|Hypothetical protein. [Gibberella zeae (Fusarium graminearum)]|('377','400')

None

>P14737@1|Q5E9Y9|58.255|0.000125|12.0|protein expressed in).[Bos taurus (Bovine)]|(794,856)|7|85|('26','82')
NDK|241-374|

>P14737@1|Q3UMN7|169.499|4.56e-08|1.0|sequence.[Mus musculus (Mouse)]|(784,815)|7|267|('96','124')
SMN|23-288|X

>P14737@1|Q5VYJ3|435.149|9.35e-05|2.3|Jumonji domain containing 2C.[Homo sapiens (Human)]|(793,831)|7|23|('949','987')
JmjN|17-62| JmjC|177-293|

>P14737@1|Q549F9|155.93|1.83e-08|1.1|inserts sequence).[Mus musculus (Mouse)]|(784,815)|7|255|('96','124')
SMN|23-288|X

>P14737@1|Q98SU9|96.913|1.99e-08|0.82|Survival motor neuron protein.[Gallus gallus (Chicken)]|(784,815)|7|238|('87','115')
SMN|12-264|X

>P14737@1|Q7KRH9|221.372|0.000282|2.4|CG5580-PB, isoform B.[Drosophila melanogaster (Fruitfly)]|('916','934')
None

>P14737@1|O02771|306.511|3.72e-09|0.7|Survival motor neuron protein.[Canis familiaris (Dog)]|(784,827)|7|264|('94','136')
SMN|21-287|X

>P14737@1|Q8T9H1|221.372|9.61e-05|3.1|SD01229p.[Drosophila melanogaster (Fruitfly)]|(838,856)|7|24|('532','550')
None

>P14737@1|Q6P684|143.448|7.81e-09|0.82|Survival of motor neuron 1, telomeric.[Rattus norvegicus (Rat)]|(784,815)|7|256|('96','124')
SMN|23-288|X

>P14737@1|O93420|157.237|7.74e-10|0.71|Survival motor neuron protein.[Brachydanio rerio (Zebrafish) (Danio rerio)]|(783,815)|7|269|('84','116')
SMN|15-285|X

>P14737@1|Q4T7G7|597.499|0.00018|24.0|(Fragment).[Tetraodon nigroviridis (Green puffer)]|(813,842)|7|1144|('449','479')
AMP-binding|281-502|X

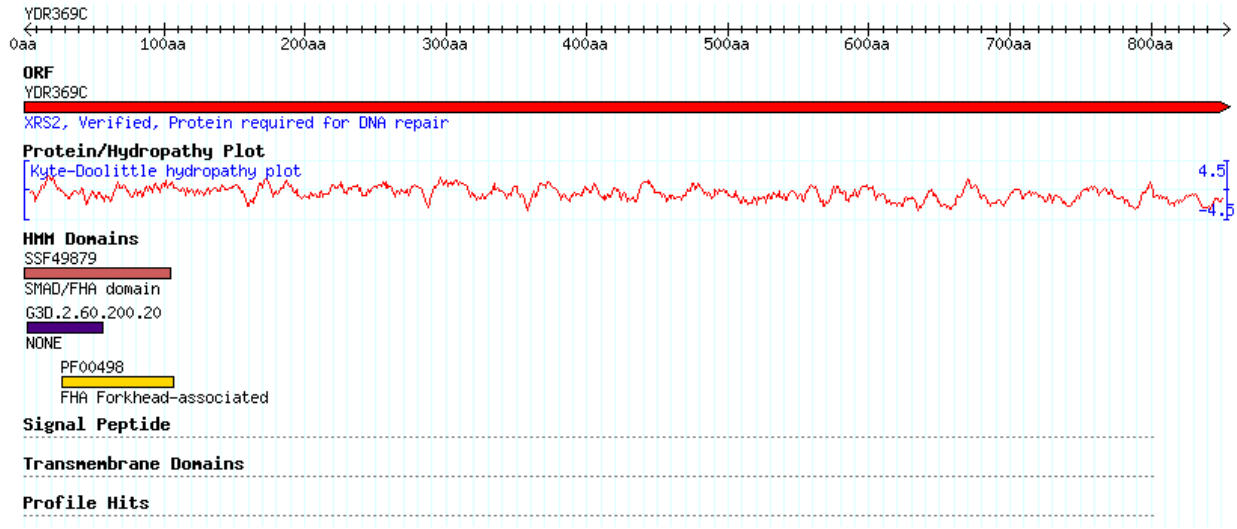
>P14737@1|Q9H3R0|485.005|0.000187|2.3|carcinoma 1 protein) (GASC-1 protein).[Homo sapiens (Human)]|(793,831)|7|23|('949','987')
JmjN|17-62| JmjC|177-293|

>P14737@1|Q9N693|221.372|9.61e-05|3.1|BrakelessB (Scribblerlongisoform).
[Drosophilamelanogaster(Fruitfly)]|(838,856)|7|24|('906','924')
None

>P14737@1|Q2TTC2|112.618|0.00027|5.7|Lipidbindingproteinag-50.[Ascaridia
galli]|('390','426')
None

XRS2
77 séquences

Légende des couleurs page 171



>P33301|2|Q9N6D4|344.293|2.36e-06|1.2|Hypothetical protein.[Leishmania major]|('1584','1623')longueur2
None

>P33301|2|Q8NHX3|83.3506|3.35e-05|2.7|Angrgm-52.[Homo sapiens (Human)]|(139,193)|6|2|('129','183')longueur2
Fusion_gly|22-241|X

>P33301|2|Q676B0|396.762|0.000117|13.0|Histone deacetylase 7A-like protein.[Oikopleura dioica]|(140,190)|6|2|('7',55')
Hist_deacetyl|237-558|

>P33301|2|O16772|288.956|7.33e-05|4.2|Hypothetical protein.[Caenorhabditis elegans]|('133','173')longueur2
None

>P33301|2|Q4P3H0|449.661|0.000311|22.0|Hypothetical protein.[Ustilago maydis 521]|(115,161)|6|606|('1395','1436')
LIM|729-773|RhoGAP|1225-1382|

>P33301|2|Q2TXC7|0.453034|1.1e-07|0.0029|Predicted protein.[Aspergillus oryzae]|('135','243')
None

>P33301|2|Q4INY1|284.174|0.000574|21.0|Predicted protein.[Gibberella zeae (Fusarium graminearum)]|('189','249')
None

>P33301|2|Q8NEM0|355.978|1.87e-07|0.0099|Microcephalin.[Homo sapiens (Human)]|(108,157)|6|460|('2',54')
BRCT|752-820|

>P33301@2|Q59FH1|277.149|4.47e-05|2.1|(Fragment).[Homo sapiens (Human)]|(125,150)|6|9|('3093','3118')
FAT|2570-2914| PI3_PI4_kinase|3267-3414|

>P33301@2|Q5TM68|324.76|2.97e-09|3.6|Mediator of DNA damage checkpoint protein 1.[Macaca mulatta (Rhesus macaque)]|(147,200)|6|508|('2003','2056')
FHA|54-124|BRCT|2075-2112|

>P33301@2|Q80YV3|277.149|4.47e-05|2.1|Transformation/transcription domain-associated protein (Tra1 homolog).[Mus musculus (Mouse)]|(125,150)|6|9|('2071','2096')
FAT|1548-1892| PI3_PI4_kinase|2245-2392|

>P33301@2|Q7R3L7|59.6984|1.14e-05|1.0|GLP_39_87490_85673.[Giardia lamblia ATCC 50803]|('330','401')
None

>P33301@2|Q515R5|344.293|0.000197|4.7|Hypothetical protein.[Entamoeba histolytica HM-1:IMSS]|(127,193)|6|2|('418','498')
LRR_1|172-192|

>P33301@2|Q5B7V8|16.7946|1.12e-11|7e-05|Hypothetical protein.[Aspergillus nidulans FGSC A4]|('23','106')
None

>P33301@2|Q5TRL2|210.442|0.000326|1.3|ENSANGP00000029084 (Fragment).[Anopheles gambiae str. PEST]|(126,160)|6|10|('3319','3353')
HEAT|1378-1414| FAT|2815-3158| PI3_PI4_kinase|3487-3638| FATC|3789-3809|

>P33301@2|Q4MYL5|254.963|0.000166|4.4|Hypothetical protein.[Theileria parva]|(167,197)|6|3|('629','659')
NIF|293-574|

>P33301@2|Q8GX29|147|0.000615|6.3|Hypothetical protein At1g31350/T19E23_7 (At1g31350).[Arabidopsis thaliana (Mouse-ear cress)]|(171,217)|6|560|('34','76')
F-box|34-79|X Kelch_2|134-181|

>P33301@2|Q4Z3Z8|390.196|0.000867|6.2|Hypothetical protein.[Plasmodium berghei]|('8','62')
None

>P33301@2|Q6FPF7|6.01812|0.000809|7.8|Similar to sp|P43579 Saccharomyces cerevisiae YFL013c.[Candida glabrata (Yeast) (Torulopsis glabrata)]|('317','403')
None

>P33301@2|Q10187|492.884|0.000392|5.1|UBA-domain containing protein 6.[Schizosaccharomyces pombe (Fission yeast)]|(158,210)|6|2|('379','431')
UBA|3-42|

>P33301@2|Q4UBK3|68.2247|6.41e-05|0.8|Hypothetical protein.[Theileria annulata]|('642','730')
None

>P33301@2|Q86ZN0|15.8418|5.58e-11|0.00025|Aspergillus nidulans.[Podospora anserina]|(129,205)|6|106|('141','220')
FHA|31-111|BRCT|141-195|

>P33301@2|Q4WAH0|0.798985|4.66e-19|3.5e-11|Camptothecin resistance conferring protein rcaA.[Aspergillus fumigatus (Sartorya fumigata)]|('30','145')
None

>P33301@2|Q9FN44|208.694|2.07e-05|0.2|Genomic DNA, chromosome 5, P1 clone:MYJ24.[Arabidopsis thaliana (Mouse-ear cress)]|(124,143)|6|11|('2639', '2658')
zf-C3HC4|4660-4693|

>P33301@2|Q5CS25|22.4904|3.38e-05|0.73|Hypothetical protein.[Cryptosporidium parvum]|('965','1020')
None

>P33301@2|Q5CJT6|75.4093|0.000298|0.75|PAN domain protein.[Cryptosporidium hominis]|(128,187)|6|4|('982','1039')
PAN_1|1359-1433|

>P33301@2|Q3TBC0|6.99337|1.47e-05|2.6|sequence.(Fragment).[Mus musculus (Mouse)]|('585','621')
None

>P33301@2|Q7PT62|184.143|2.96e-09|0.0066|ENSANGP00000017923.[Anopheles gambiae str.PEST]|(132,181)|6|537|('228','272')
BRCT|202-279|X

>P33301@2|Q5CXZ2|29.6194|3.48e-05|0.97|Very large secreted protein, signal peptide.[Cryptosporidium parvum]|('872','930')
None

>P33301@2|Q14676|424.15|5.51e-09|0.095|domains 1).[Homo sapiens (Human)]|(147,200)|6|559|('1919','1972')
FHA|54-124|BRCT|1991-2028|

>P33301@2|Q4N9N6|77.9686|1.71e-05|0.73|Hypothetical protein.[Theileria parva]|(170,216)|6|3|('620','672')
HECT|1389-1725|

>P33301@2|Q7YYR5|86.9015|0.000298|0.75|PAN domain protein.[Cryptosporidium parvum]|(128,187)|6|4|('969','1026')
PAN_1|1364-1438|TRAPP_Bet3|2278-2324|

>P33301@2|Q4S4H6|277.149|2.49e-05|2.0|Chromosome 2 SCAF14738, whole genome shotgun sequence.[Tetraodon nigroviridis (Green puffer)]|('2659','2684')
None

>P33301@2|Q7PUY7|73.5452|3.74e-10|0.0043|ENSANGP00000011531 (Fragment).[Anopheles gambiae str.PEST]|(132,187)|6|562|('47','97')
BRCT|291-329|

>P33301@2|Q4SYS4|187.242|7.32e-10|0.0038|(Fragment).[Tetraodon nigroviridis (Green puffer)]|(130,183)|6|321|('37','101')
BRCT|14-91|X AAA|267-480| RFC1|571-724|

>P33301@2|Q8I7C6|406.819|5.87e-10|0.0029|adprt2).[Dictyostelium discoideum (Slime mold)]|(135,186)|6|574|('56','106')
BRCT|29-108|X WGR|233-319| PARP_reg|360-491| PARP|493-700|

>P33301@2|Q55TB1|179.591|5.25e-05|2.4|Hypothetical protein.[Cryptococcus neoformans var.neoformans B-3501A]|('256','294')
None

>P33301@2|Q7PLT6|257.099|3.4e-13|0.0001|CG40411-PC.3 (RE04933p). [Drosophila melanogaster (Fruit fly)]|(129,184)|6|184|('397','454')
zf-PARP|114-197| PADR1|269-323| BRCT|380-458|X WGR|533-616|
PARP_reg|644-777| PARP|779-990|

>P33301@2|Q767L8|166.599|6.29e-10|3.0|Mediator of DNA damage checkpoint protein 1. [Sus scrofa (Pig)]|(147,200)|6|549|('1872','1925')
FHA|54-124|BRCT|1944-2022|

>P33301@2|Q2R2B4|45.3297|1.79e-09|0.0068|BRCA1 C Terminus domain, putative.[Oryza sativa (japonica cultivar-group)]|(129,207)|6|365|('266','344')
BRCT|248-325|X AAA|452-564| RFC1|726-889|

>P33301@2|Q2TAZ4|424.15|5.51e-09|0.095|Hypothetical protein (Fragment).[Homo sapiens (Human)]|(147,200)|6|559|('618','671')
BRCT|690-727|

>P33301@2|Q8N0N1|301.266|0.000749|48.0|SHG. [Littorina littorea (Common periwinkle)]|('103','158')
None

>P33301@2|Q4N0Z0|108.859|5.49e-05|0.71|Hypothetical protein.[Theileria parva]|('392','436')
None

>P33301@2|Q7S501|2.84009|1.28e-10|0.00017|Hypothetical protein.[Neurospora crassa]|('39','90')
None

>P33301@2|Q7S0D7|279.471|0.00015|3.1|Predictedprotein.[Neurosporacrassa]|('998','1051')
None

>P33301@2|Q7R5Q7|57.2591|1.56e-05|8.1|GLP_487_25078_22760.[Giardia lambliaATCC 50803]|('383','424')
None

>P33301@2|Q9GFL3|513.882|0.000723|15.0|Chloroplast30S ribosomalproteinS7.[Ginkgobiloba(Ginkgo)]|(174,213)|6|354|('13','52')
Ribosomal_S7|1-149|X

>P33301@2|Q3UH32|277.149|4.47e-05|2.1|homolog)homolog(Fragment).[Mus musculus(Mouse)]|(125,150)|6|9|('2230','2255')
FAT|1707-2051| PI3_PI4_kinase|2404-2551|

>P33301@2|Q5KA92|6.2745|9.8e-06|1.7|Expressedprotein.[Cryptococcus neoformans(Filobasidiellaneoformans)]|('1','105')
None

>P33301@2|Q9Y4A5|277.149|4.47e-05|2.1|PCAF-associatedfactor)(PAF350/400)(STAF40)(Tra1homolog).[Homo sapiens(Human)]|(125,150)|6|9|('3365','3390')
FAT|2849-3204| PI3_PI4_kinase|3539-3686|

>P33301@2|Q55L91|6.2745|9.8e-06|1.7|Hypotheticalprotein.[Cryptococcus neoformansvar.neoformansB-3501A]|('1','105')
None

>P33301@2|Q5LJS3|257.099|3.4e-13|0.0001|CG40411-PD.3.[Drosophila melanogaster(Fruitfly)]|(129,184)|6|184|('397','454')
zf-PARP|114-197| PADR1|269-323| BRCT|380-458|X WGR|533-615|

>P33301@2|Q4FG85|513.882|0.000723|15.0|RibosomalproteinS7 (Fragment).[Ginkgobiloba(Ginkgo)]|(174,213)|6|354|('13','52')
Ribosomal_S7|1-149|X

>P33301@2|Q4UG38|420.626|1.71e-05|0.73|Ubiquitin-relatedprotein,putative.[Theileriaannulata]|(170,216)|6|3|('633','685')
HECT|1400-1736|

>P33301@2|Q3V061|92.9001|0.000127|0.12|fullinsertsequence.[Mus musculus(Mouse)]|(114,203)|6|3|('383','448')
zf-C3HC4|71-114| zf-B_box|206-246| SPRY|497-624|

>P33301@2|Q7YR40|424.15|2.18e-09|0.055|MediatorofDNA damage checkpoint protein1.[Pan troglodytes(Chimpanzee)]|(147,200)|6|540|('2001','2054')
FHA|54-124|BRCT|2073-2110|

>P33301@2|Q9SHE6|147|0.000903|14.0|T19E23.14.[Arabidopsisthaliana(Mouse-earcress)]|(171,217)|6|477|('85','127')

F-box|85-130|X Kelch_2|185-232|

>P33301@2|Q54HY5|15.0682|6.98e-10|0.014|Hypothetical protein.[Dictyostelium discoideum (Slime mold)]|(134,202)|6|498|('24','91')
BRCT|1-77|X WGR|271-350| PARP|505-688| Ank|1415-1444|

>P33301@2|Q501K4|14.8188|7.12e-06|3.9|Hypothetical protein (Fragment). [Xenopus tropicalis (Western clawed frog) (Silurana tropicalis)]|(108,198)|6|2|('270', '366')
V-set|36-147| ig|232-309|X

>P33301@2|Q6C7A1|0.11822|5.19e-11|6.4e-05|Similarity.[Yarrowia lipolytica (Candida lipolytica)]|('116','171')
None

>P33301@2|Q5IFK1|355.978|1.87e-07|0.0081|Microcephalin.[Macaca fascicularis (Crab eating macaque) (Cynomolgus monkey)]|(108,157)|6|466|('2',54')
BRCT|759-827|

>P33301@2|P35875|257.099|3.4e-13|0.0001|ribosyltransferase) (Poly[ADP-ribose] synthetase).[Drosophila melanogaster (Fruit fly)]|(129,184)|6|184|('397','454')
zf-PARP|114-197| PADR1|269-323| BRCT|380-458|X WGR|533-616|
  PARP_reg|644-777| PARP|779-990|

>P33301@2|Q7Q731|210.442|0.000326|1.3|ENSANGP00000007163.[Anopheles gambiae str.PEST]|(126,160)|6|10|('2951','2985')
HEAT|1030-1066| FAT|2418-2761| PI3_PI4_kinase|3119-3270| FATC|3431-3451|

>P33301@2|Q9P4A3|16.7946|1.12e-11|7e-05|Camptothecin resistance conferring protein rcaA.[Emericella nidulans (Aspergillus nidulans)]|(126,205)|6|89|('23','106')
None

>P33301@2|Q84N08|45.3297|1.79e-09|0.0068|Replication factor C 110 kDa subunit. [Oryza sativa (japonica cultivar-group)]|(129,207)|6|365|('275','353')
BRCT|257-334|X AAA|461-573| RFC1|735-898|

>P33301@2|Q6PYB5|263.616|9.7e-08|0.0052|Microcephalin.[Saimiri boliviensis (Bolivian squirrel monkey)]|(108,157)|6|433|('2',54')
BRCT|758-826|

>P33301@2|Q6PYB7|449.661|1.06e-07|0.0067|Microcephalin.[Canis familiaris (Dog)]|(107,157)|6|464|('7',60')
BRCT|778-835|

>P33301@2|Q5CWU1|86.9015|0.00029|0.75|transmembrane region. [Cryptosporidium parvum]|(128,187)|6|4|('982','1039')
PAN_1|1377-1451|

>P33301@2|Q5KJ28|179.591|5.25e-05|2.4|Hypothetical protein.[Cryptococcus neoformans (Filobasidiella neoformans)]|('256','294')
None

>P33301@2|Q3U0G9|6.99337|1.47e-05|2.6|(Fragment).[Mus musculus (Mouse)]|('585','621')
None

>P33301@2|Q94HC1|14.3324|4.15e-06|11.0|repeat, putative).[Oryza sativa (japonica cultivar-group)]|(131,182)|6|4|('218','266')
PUF|586-620|

>P33301@2|Q51GR8|464.921|0.000179|2.9|Hypothetical protein.[Entamoeba histolytica HM-1:IMSS]|('15','62')
None

>P33301@2|Q6BFX8|296.28|5.13e-05|2.5|Hypothetical protein.[Paramecium tetraurelia]|('467','508')
None

>P33301@2|Q9TX05|488.789|2.2e-05|3.0|NAD+:PROTEIN (ADP-ribosyl)-transferase, ADPRT (Fragment).[Drosophila sp. (Fruit fly)]|(132,184)|6|27|('2','53')
WGR|132-215| PARP_reg|243-376| PARP|378-589|

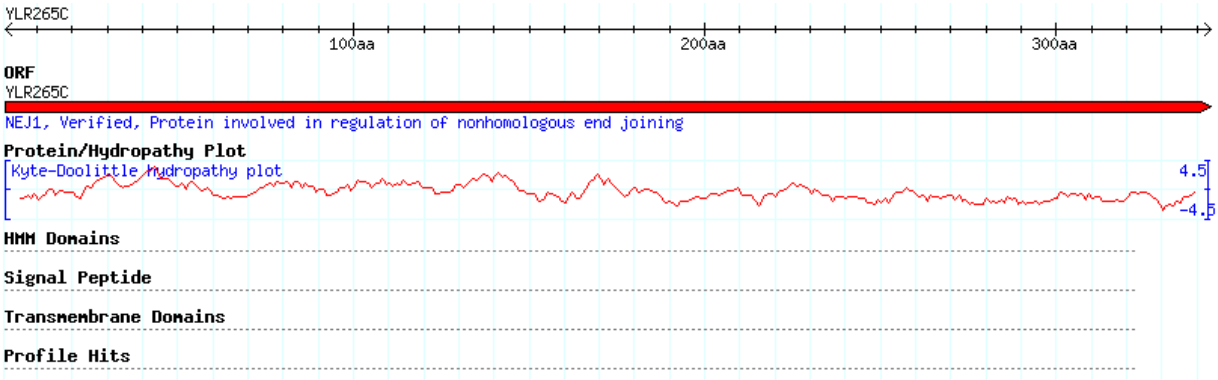
>P33301@2|P61594|355.978|3.54e-07|0.0062|Microcephalin.[Pongo pygmaeus (Orangutan)]|(108,157)|6|447|('2','54')
BRCT|756-824|

>P33301@2|P61591|355.978|3.54e-07|0.0062|Microcephalin.[Gorilla gorilla gorilla (Lowland gorilla)]|(108,157)|6|447|('2','54')
BRCT|752-820|

>P33301@2|P61590|355.978|3.54e-07|0.0062|Microcephalin.[Colobus guereza (Black-and-white colobus monkey)]|(108,157)|6|447|('2','54')
BRCT|758-826|

NEJ1
34 séquences

Légende des couleurs page 171



>Q06148@1|Q9LGA5|232.279|0.00094|2.6|Putativenuclease I.[Oryza sativa (japonicacultivargroup)]|(188,262)|4|75|('167','234')
S1-P1_nuclease|34-294|X

>Q06148@1|Q4UDJ5|64.802|0.00014|6.9|Pyruvatekinase,putative(EC 2.7.1.40). [Theileriaannulata]|(22,85)|4|3|('588','653')
PK|116-543|

>Q06148@1|Q9ZR89|213.685|0.00014|0.54|Bifunctionalnucleasebfn1. [Arabidopsisthaliana(Mouse-earcress)]|(188,258)|4|72|('167','233')
S1-P1_nuclease|29-294|X

>Q06148@1|O13515|12.214|0.00022|5.7|Yal035cap (YAL035C-A). [Saccharomyces cerevisiae(Baker'syeast)]|('41','117')
None

>Q06148@1|Q7YY84|364.492|0.00029|0.31|Hypotheticalprotein.[Cryptosporidium parvum]|('456','485')
None

>Q06148@1|Q5CKJ7|182.358|6.08e-06|2.3|Hypotheticalprotein.[Cryptosporidium hominis]|('487','577')
None

>Q06148@1|Q510L7|409.655|0.00029|8.1|Ubiquitincarboxylterminalhydrolase, putative.[Entamoeba histolyticaHM-1:IMSS]|(113,203)|4|2|('2032','2103')
UCH|1350-1646|

>Q06148@1|Q4D889|174.907|0.00033|56.0|Hypotheticalprotein(Fragment). [Trypanosoma cruzi]|('1365','1509')
None

>Q06148@1|Q8IBG1|349.598|0.00041|0.87|Dyneinheavy chain,putative. [Plasmodium falciparum (isolate3D7)]|(65,149)|4|271|('2152','2243')

DHC_N1|280-856| DHC_N2|1401-1810| AAA_5|2314-2465| Dynein_heavy| 4902-4970|

>Q06148@1|Q6KAU8|305.908|4.94e-05|1.8|MFLJ00012 protein(Fragment).[Mus musculus(Mouse)]|(211,232)|4|15|('26','47')
ATG16|27-209|X WD40|400-437|

>Q06148@1|Q55BX4|98.3502|0.00015|4.7|Hypothetical protein.[Dictyostelium discoideum (Slime mold)]|(131,185)|4|2|('6','63')
LRR_1|601-623|

>Q06148@1|Q4UAU7|24.2105|9.45e-06|3.6|Hypothetical protein.[Theileria annulata]|('180','238')
None

>Q06148@1|Q9VIK7|232.279|5.71e-07|0.073|CG31678-PA. [Drosophila melanogaster(Fruitfly)]|(135,161)|4|3|('6';32')
Sad1_UNC|613-736|

>Q06148@1|Q6PAU0|305.908|0.00030|9.7|Atg16l2 protein(Fragment).[Mus musculus(Mouse)]|(211,232)|4|16|('90','111')
ATG16|91-286|X WD40|404-442|

>Q06148@1|Q5CGS8|193.326|0.00024|7.5|Hypothetical protein.[Cryptosporidium hominis]|('166','226')
None

>Q06148@1|Q8T1Q2|327.024|0.00022|5.7|Tcc44h21-2.7.[Trypanosoma cruzi]|('808','884')
None

>Q06148@1|Q5CLE1|364.492|0.00029|0.31|Hypothetical protein.[Cryptosporidium hominis]|('456','485')
None

>Q06148@1|Q9Y485|13.7276|8.72e-05|19.0|X-like1 protein.[Homo sapiens (Human)]|(57,186)|4|1|('2590','2702')
WD40|2918-2956|

>Q06148@1|Q7QTI2|308.471|0.00021|5.7|GLP_251_10678_7133.[Giardialamblia ATCC 50803]|('743','783')
None

>Q06148@1|Q6MWN9|185.427|3.53e-05|45.0|Hypothetical protein B10K17.100. [Neurosporacrassa]|('336','403')
None

>Q06148@1|Q7YU09|298.346|5.71e-07|0.073|LD18032p.[Drosophilamelanogaster (Fruitfly)]|(135,161)|4|3|('6';32')
Sad1_UNC|409-532|

>Q06148@1|Q9XEF3|252.492|0.000106|2.9|Hypothetical cytoskeletal protein. [Arabidopsis thaliana (Mouse-ear cress)]|(162,266)|4|3|('221','323') Myb_DNA-binding|84-135|

>Q06148@1|Q7QXM2|15.5579|0.00076|5.7|GLP_512_53165_50094. [Giardia lamblia ATCC 50803]|('665','764') None

>Q06148@1|Q95WV2|361.463|4.97e-05|47.0|Collagen. [Meloidogyne javanica (Root-knot nematode worm)]|(152,209)|4|501|('37','104') Col_cuticle_N|44-86|X Collagen|286-321|

>Q06148@1|Q4YZY5|441.602|0.00091|4.0|Hypothetical protein (Fragment). [Plasmodium berghei]|('229','260') None

>Q06148@1|Q7SC11|185.427|3.53e-05|45.0|Predicted protein. [Neurospora crassa]|('389','456') None

>Q06148@1|Q4D8Q7|172.012|0.000515|16.0|Hypothetical protein. [Trypanosoma cruzi]|('1329','1396') None

>Q06148@1|Q7R1V6|22.2724|7.19e-05|2.0|GLP_190_1747_6126. [Giardia lamblia ATCC 50803]|('1223','1307') None

>Q06148@1|Q9SXA6|213.685|0.000145|0.54|Bifunctional nuclease bfn1. [Arabidopsis thaliana (Mouse-ear cress)]|(188,258)|4|72|('167','233') S1-P1_nuclease|29-294|X

>Q06148@1|Q4E5J0|358.46|0.000225|7.1|Hypothetical protein. [Trypanosoma cruzi]|('808','884') None

>Q06148@1|Q9SIZ5|228.435|0.000109|2.9|Hypothetical protein At2g40260. [Arabidopsis thaliana (Mouse-ear cress)]|(162,266)|4|3|('221','323') Myb_DNA-binding|84-135|

>Q06148@1|Q7QY84|18.8492|2.69e-06|26.0|GLP_572_11666_9606. [Giardia lamblia ATCC 50803]|('76','208') None

>Q06148@1|Q2QTF0|186.98|1.58e-05|0.74|Retrotransposon protein, putative, Ty3-gypsy sub-class. [Oryza sativa (japonica cultivar-group)]|('50','99') None

>Q06148@1|Q5CV12|113.338|6.08e-06|1.5|transferase,transcriptsidentifiedby
EST.[Cryptosporidiumparvum]|('487','577')
None

VIII.2.Article 1

BIOINFORMATICS DISCOVERY NOTE Vol. 22 no. 11 2006, pages 1289–1292
doi:10.1093/bioinformatics/btl075

Structural bioinformatics

Detection of a tandem BRCT in Nbs1 and Xrs2 with functional implications in the DNA damage response

Emmanuelle Becker[1,†], Vincent Meyer[2,†], Hocine Madaoui[1] and Raphaël Guerois[1,*]

[1]Service de Biophysique des Fonctions Membranaires, URA CNRS 2096, Département de Biologie Joliot-Curie and [2]Département d'Etude et d'Ingénierie des Protéines, CEA Saclay, 91191 Gif-Sur-Yvette, Cedex, France

Received on January 30, 2006; revised and accepted on February 27, 2006
Advance Access publication March 7, 2006
Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Human Nbs1 and its homolog Xrs2 in *Saccharomyces cerevisiae* are part of the conserved MRN complex (MRX in yeast) which plays a crucial role in maintaining genomic stability. NBS1 corresponds to the gene mutated in the Nijmegen breakage syndrome (NBS) known as a radiation hyper-sensitive disease. Despite the conservation and the importance of the MRN complex, the high sequence divergence between Nbs1 and Xrs2 precluded the identification of common domains downstream of the N-terminal Fork-Head Associated (FHA) domain.

Results: Using HMM–HMM profile comparisons and structure modelling, we assessed the existence of a tandem BRCT in both Nbs1 and Xrs2 after the FHA. The structure-based conservation analysis of the tandem BRCT in Nbs1 supports its function as a phosphoserine binding domain. Remarkably, the 5 bp deletion observed in 95% of NBS patients cleaves the tandem at the linker region while preserving the structural integrity of each BRCT domain in the resulting truncated gene products.

Contact: guerois@cea.fr

Supplementary information: http://www-spider.cea.fr/Groups/si6661/view.html

1 INTRODUCTION

Nbs1 in human (or Xrs2 in yeast) is an essential component of the so-called MRN complex associating Mre11, Rad50 and Nbs1 (Petrini and Stracker, 2003; van den Bosch *et al.*, 2003) and plays a crucial role in DNA repair pathways (Kobayashi *et al.*, 2004). The human Nbs1 protein is a 754 amino acid long protein composed of several functional domains identified from sequence analysis and biochemical experiments (Fig. 2A). At the N-terminus, a Fork-Head Associated (FHA) domain (Durocher and Jackson, 2002) followed by a single BRCA1 C-terminal (BRCT) domain (Bork *et al.*, 1997; Callebaut and Mornon, 1997) can be detected from sequence to profile searches. The C-terminus of Nbs1 contains a Mre11 binding region (Desai-Mehta *et al.*, 2001) and an ATM recruitment motif (Falck *et al.*, 2005). In Xrs2, the *Saccharomyces cerevisiae* functional homolog of Nbs1, the FHA domain together with the Tel1 (ATM homologue) and Mre11 binding regions are conserved but the existence of a BRCT domain was never detected from sequence analysis. As a matter of fact, the sequences of Xrs2 and Nbs1 are highly divergent in the 250 amino acids following the FHA domain (10% sequence identity). Using a specific strategy, new sequences of Xrs2 homologs not present in databases such as GenBank or EMBL could be retrieved and aligned to human sequences. From the resulting multiple sequence alignment, we show that in fact two BRCT domains are present in both human Nbs1 and yeast Xrs2 right behind the FHA domain.

Tandem BRCT have been recently recognized as major mediators of phosphorylation-dependent protein–protein interactions in processes related to cell-cycle checkpoint and DNA repair functions (Glover *et al.*, 2004). The ability of the tandem BRCT of Nbs1 to bind phospho-peptides was never probed before since the existence of the second BRCT was not suspected. The model-based analysis of the tandem BRCT of Nbs1 strongly suggests that it is a phosphoserine binding module. The 5 bp deletion observed in 95% of NBS patients splits up the tandem at position 218. Remarkably, this mutation preserves the structural integrity of the second BRCT at plus or minus one residue. Altogether, our findings suggest that the NBS disease could be partly linked to a disruption of the interaction properties of the tandem BRCT: cleavage of the tandem BRCT may alter the selectivity of target recognition by Nbs1 and hence affect the signaling network required for efficient DNA damage responses.

2 METHODS

An initial profile containing close homologs of Nbs1 was built from searches of the non-redundant database using PSI-BLAST (Altschul *et al.*, 1997) on the MPI server (Soding *et al.*, 2005). For Xrs2, the initial profile gathered three sequences retrieved from tblastn searches on the *Saccharomyces* comparative genomic database (Kellis *et al.*, 2003). The profiles were enriched by aligning profiles of more divergent sequences using the profile–profile alignment method HHalign (Soding, 2005).

Iteratively, the profile–profile alignment procedure led to a global multiple sequence alignment gathering 25 sequences from human Nbs1 to *S.cerevisiae* Xrs2 (see Supplementary information). The profile consisting of 25 sequences was compared against a database of profiles built from the PDB using the HHpred server (Soding *et al.*, 2005). Three structures of

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

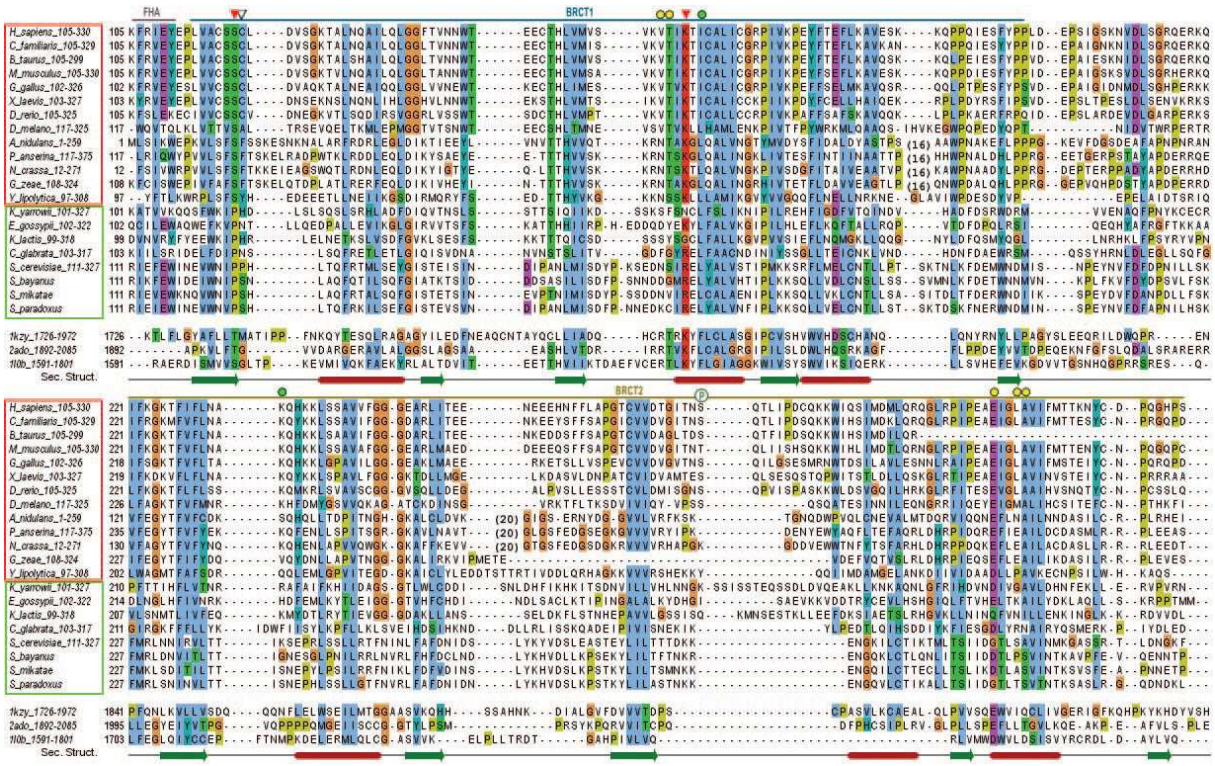


Fig. 1. Multiple sequence alignment of the tandem BRCT in homologs of Nbs1 and Xrs2 aligned with three structures of tandem BRCT (1kzy: 53BP1, 2ado: MDC1, 1l0b: BRCA1) represented with Jalview (Clamp *et al.*, 2004). Helices and strands are noted by red sticks and green arrows below. Domains boundaries are shown by an horizontal line above. Red and green boxes group the sequences names with respect to their patterns at the pSer binding positions. Positions contacting the pSer residue with sidechain or backbone in tandem BRCT complex are indicated by red and white triangles, respectively. Positions contacting the pSer+3 positions are shown by green circles. Other positions found in direct contact with the phosphopeptide atoms are shown by yellow circles.

tandem BRCT were detected with significant scores and confidence levels >96% (PDB codes 1l0b, 2ado, 1kzy). Over the major length of the profile, the built alignment was consistent with the structural alignment of the templates. Yet, significant divergence could be observed at the N-terminus (first strand) and C-terminus (last $\alpha/\beta/\alpha$ motif). At the N-terminus, only the alignments with 1l0b and 2ado were compatible with the presence of an upstream FHA domain. At the C-terminus, only the alignment with the 1kzy template suggested the existence of a long insertion in the $\beta3\alpha2$ loop of the second BRCT, consistent with the conservation profile in the whole Nbs1 family. A global sequence to structure alignment between the 25 sequences and the structural alignment of the three templates (1l0b, 2ado and 1kzy) was created based on these features.

Models were generated for both human Nbs1 and *S.cerevisiae* Xrs2 with Modeller 8v2 (Sali and Blundell, 1993) using the three structures 1kzy, 2ado and 1l0b as templates (max. Seq. ID: 13.2%). The quality of the models was assessed using Verify3D (Luthy *et al.*, 1992), Prosa2003 (Sippl, 1993), ProQ and MaxSub (Wallner and Elofsson, 2003). The profile–profile alignment between the tandem BRCT of the Nbs1/Xrs2 family and that of the structural alignment of the three templates was iteratively refined in order to reduce the alignment errors pinpointed by the four evaluation scores.

To further assess the physical relevance of the model built for the tandem BRCT of Nbs1, a 5 ns molecular dynamic simulation was performed at 300 K in explicit solvent using GROMACS 3.2 (Van Der Spoel *et al.*, 2005) (see Supplementary information for details). Conservation analyses were carried out using the Rate4site algorithm (Pupko *et al.*, 2002). Possible arrangements of the FHA domain with respect to the tandem BRCT were explored using the HADDOCK program (Dominguez *et al.*, 2003) by docking models of the FHA domain onto models of the tandem BRCT while constraining the distance between their C- and N-termini in respect of the Nbs1 sequence.

3 RESULTS

Models of the Nbs1 and Xrs2 tandem BRCT were built from the multiple sequence alignment in Figure 1 and assessed using standard evaluation tools. The scores of Nbs1 model are (Prosa2003: −1.92), (Verify3D: 0.395), (ProQ : 3.51) and (MaxSub : 0.348) and those of Xrs2 (Prosa2003: −1.16), (Verify3D: 0.332), (ProQ : 3.75) and (MaxSub : 0.338). The absence of residues with Verify3D scores below 0.1 together with ProQ and MaxSub scores significantly above 1.5 and 0.1, respectively, ensures the absence of major issues in the models of both tandem BRCT (Wallner and Elofsson, 2003). The physical quality of the Nbs1 model was further assessed by running a 5 ns simulation of molecular dynamics in an explicit solvent. The Cα rmsd stabilizes around 4 Å from the initial model structure (3.2 Å excluding the long loops 201–216 and 273–291) and secondary structures are overall preserved after a 5 ns of simulation as illustrated in Figure 2B (see also Supplementary information).

3.1 Functional insights from the tandem BRCT model

3.1.1 Clues for phosphoserine binding in Nbs1 So far, the tandem BRCT repeats of MDC1, PTIP, BARD1, 53BP1, RAD4, Ect2, TOPBP1, DNA ligase IV, *S.pombe* Crb2 and *S.cerevisiae* Rad9 have been shown to have phospho-serine (pSer) binding properties *in vitro* (Manke *et al.*, 2003; Yu *et al.*, 2003). The consensus signature for the pSer binding property was described as [S/T-G] in $\beta1\alpha1$ loop

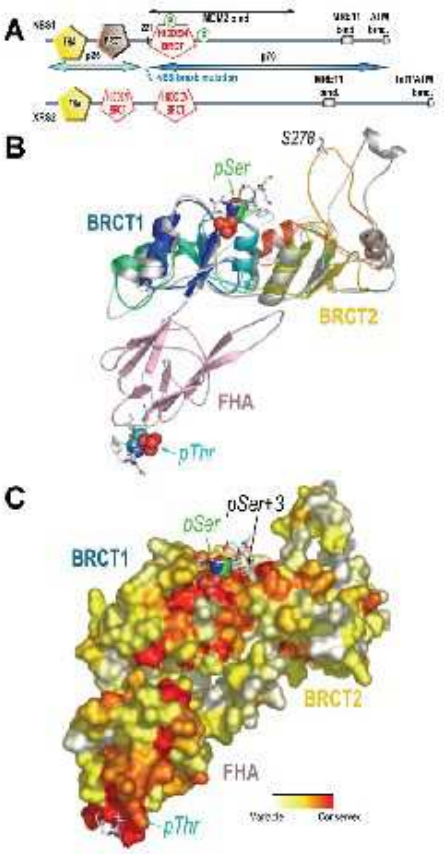


Fig. 2. (A) Domain organization of Nbs1 and Xrs2 with newly identified BRCT shown in red. (B) Ribbon representation of the model of the tandem BRCT before (rainbow colors) and after a 5 ns simulation of molecular dynamics (gray). In pink, a model of the FHA domain in a putative orientation with respect to the tandem. Phospho-peptides as found in known complexes of FHA and tandem BRCT with their ligands are shown as sticks. Phospho-Ser and -Thr residues are shown as spheres. (C) Surface projection of the evolutionary rates as calculated by the rate4site algorithm (Pupko *et al.*, 2002). Colors from red to white report from the most conserved to the most variable positions. Drawn with pymol (DeLano, 2002).

and [S/T-X-K] in α2 helix of the first BRCT (Glover *et al.*, 2004). The Gly in β1/α1 loop is quite versatile probably because only the backbone atoms at that position interacts with the pSer residue [53BP1 (1kzy in Fig. 1) has a Met and binds pSer *in vitro*]. In the tandem BRCT of Nbs1, the most conserved region localizes in the sites shown to bind the pSer residue in the structures of the tandem BRCT complexes (Stucki *et al.*, 2005). The positions that directly contact the pSer with their sidechain or backbone are indicated by red and white triangles, respectively (Fig. 1). In sequences from *Homo sapiens* Nbs1 to *Yarrowia lipolytica* fungus (red box, Fig. 1), the consensus motif [S-OF] in β1/α1 loop and [T/S-X-K] in α2 is strictly conserved supporting the function of this module as a pSer binding domain. In species ranging from *Kluyveromyces yarrowii* to *Solenodon paradoxus* (green box, Fig. 1), including Xrs2, positions binding pSer with their sidechains (red triangles) are conserved but do not match the consensus phospho-binding signature (Glover *et al.*, 2004). The corresponding motif in *S.cerevisiae* Xrs2 is [P-P] in β1/α1 loop and [S-X-R] in α2 helix. Yet, several clues support that the tandem BRCT of Xrs2 might still be a pSer-binding module: (1) the BRCT domain of the ligase III shown to bind pSer peptides *in vitro* contains a Pro instead of a Ser in β1/α1 loop, as in Xrs2, (2) an Arg in α2 instead of a Lys is found in the tandem BRCT of ligase IV, also shown to be a pSer-binding module *in vitro* (Yu *et al.*, 2003).

The groove at the interface between the BRCT domains is involved in the specific recognition of the residues flanking the pSer amino acid and is significantly conserved in the Nbs1 family (Fig. 2C). In structures of tandem BRCT/phosphopeptide complexes, the pSer+3 position was shown to hold much of the binding selectivity (Glover *et al.*, 2004). Positions whose sidechain were shown to directly contact the position pSer+3 are indicated by green circles in Figure 1. In contrast to known structures where hydrophobic residues are often found at those positions, a Lys is quite conserved in one position of the Nbs1 multiple alignment (K233 in *H.sapiens* Nbs1).

3.1.2 Location of the phosphorylated sites in Nbs1 In response to ionizing radiation, Nbs1 is phosphorylated at Ser278 and Ser343 by the ATM kinase, and this event is required for activation of the intra S phase checkpoint (Kobayashi *et al.*, 2004). From the structural model, Ser278 is located in the long β3/α2 loop of the second BRCT (Fig. 2B) and Ser343 is found 13 residues after the last residue of the tandem BRCT. Interestingly, the flexible linkers surrounding Ser278 and Ser343 are not long enough to allow for an intramolecular recognition of the pSer by the tandem BRCT.

3.1.3 Tandem BRCT and disease related mutations Of the NBS patients, 95% carry a 5 bp deletion in exon 6 of the *NBS1* gene, which results in the expression of two truncated proteins of 26 (p26) and 70 kDa (p70) (Fig. 2A). The mutation splits the tandem precisely in the linker between the two BRCT domains. P26 moiety includes the region 1–218 spanning the FHA and the integrality of the first BRCT domain. P70 corresponds to the C-terminal half of Nbs1 and is produced by an alternative initiation of translation upstream of the 5 bp deletion. After a 18 residue extension at the N-terminus, the sequence of p70 is identical to that of the wild-type Nbs1 from I221 to the end (Williams *et al.*, 2002).

I221 sharply corresponds to the beginning of the second BRCT and is the first residue fully buried in its hydrophobic core. Several structures of well-folded single C-terminal BRCT domains isolated from a tandem support that each BRCT domain can adopt its structure independently (Gaiser *et al.*, 2004; Zhang *et al.*, 1998). Hence, despite the severe sequence variations induced by the mutation in the linker, elements crucial for the structural integrity of the second BRCT have been preserved. It suggests that the second BRCT may not only fold independently but also hold a function important for

viability in NBS patients. Regarding the first BRCT, it has been shown that the FHA/BRCT could bind *in vitro* the histone H2AX phosphorylated by ATM (Kobayashi *et al*., 2002). Phosphorylation of H2AX at Ser129 is among the first events of the repair of double strand breaks (Lowndes and Toh, 2005). Our data suggest that the p26 fragment (Fig. 2A) may still be able to bind pSer residues in NBS cells but with a loss of binding selectivity due to the truncation of the second BRCT. This novel hypothesis would be interesting to test in the light of the results obtained on animal models of the NBS pathology (Difilippantonio *et al*., 2005; Williams *et al*., 2002).

3.1.4 Nbs1 and Mdm2 interaction Mdm2 has been extensively studied as a negative regulator of p53 tumor suppressor (Vousden and Prives, 2005). Mdm2 overexpression was recently shown to inhibit the DNA repair function of the MRN complex and this effect required the binding of Mdm2 to Nbs1 (Alt *et al*., 2005). The region 198–314 of Mdm2 was shown to associate with the MRN complex through the central region of Nbs1 221–540. This region encompasses the newly identified second BRCT domain 221–330 but not the first one. Downstream of the second BRCT, the region 330–540 is predicted to be largely unfolded (see Supplementary information). We hypothesize that the second BRCT of Nbs1 by itself may be involved in the interaction with Mdm2.

3.2 Functional implications from the FHA-tandem BRCT structural model

A striking feature of the domain organization among all Nbs1 homologs is the absence of a linker between the FHA and the tandem BRCT modules. Despite the high versatility in position and length of the insertions inside the FHA or the BRCT and between the two BRCT, not even a single amino acid was ever added at the hinge between the two modules. A structural model of the ensemble composed by the FHA and the tandem BRCT domains was built to probe the potential organization of the modules (Fig. 2B). Owing to steric hindrance, the phospho-binding sites of both domains are constrained on opposite sides of the whole assembly and could hardly be closer than 45 Å (see Supplementary information). It excludes the possibility to bind simultaneously a pThr neighboring a pSer at <15 residues. The structural constraint between the domains may originate from a specific evolutionary constraint coupling both pThr and pSer binding functions. Interestingly, the FHA and the BRCT were shown to be both required for optimal chromatin association of the MRN complex (Kobayashi *et al*., 2002; Zhao *et al*., 2002). Moreover, a mutation disrupting the FHA pThr binding site revealed that this domain is involved in a signal amplification step crucial for DNA repair after low doses of irradiation (Difilippantonio *et al*., 2005). The coupling between pThr and pSer binding functions suggested from the model might as well contribute to this amplification process.

ACKNOWLEDGEMENTS

The authors are grateful to F. Ochsenbein, M.-C. Marsolier-Kergoat and S. Zinn-Justin for their useful comments about the manuscript. This work is partly funded by the ACI IMPBIO 2004. V.M. is supported by an AFM fellowship (Association Française contre les Myopathies). H.M. is supported by a DGA fellowship. Funding to pay the Open Access publication charges was provided by the CEA Saclay.

Conflict of Interest: none declared.

REFERENCES

Alt,J.R. *et al.* (2005) Mdm2 binds to Nbs1 at sites of DNA damage and regulates double strand break repair. *J. Biol. Chem.*, 280, 18771–18781.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25, 3389–3402.

Bork,P. *et al.* (1997) A superfamily of conserved domains in DNA damage-responsive cell cycle checkpoint proteins. *FASEB J.*, 11, 68–76.

Callebaut,I. and Mornon,J.P. (1997) From BRCA1 to RAP1: a widespread BRCT module closely associated with DNA repair. *FEBS Lett.*, 400, 25–30.

Clamp,M. *et al.* (2004) The Jalview Java alignment editor. *Bioinformatics*, 20, 426–427.

DeLano,W.L. (2002) *The PyMOL Molecular Graphics System*. DeLano Scientific, San Carlos, CA, USA.

Desai-Mehta,A. *et al.* (2001) Distinct functional domains of nibrin mediate Mre11 binding, focus formation, and nuclear localization. *Mol. Cell. Biol.*, 21, 2184–2191.

Difilippantonio,S. *et al.* (2005) Role of Nbs1 in the activation of the Atm kinase revealed in humanized mouse models. *Nat. Cell Biol.*, 7, 675–685.

Dominguez,C. *et al.* (2003) HADDOCK: a protein–protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.*, 125, 1731–1737.

Durocher,D. and Jackson,S.P. (2002) The FHA domain. *FEBS Lett.*, 513, 58–66.

Falck,J. *et al.* (2005) Conserved modes of recruitment of ATM, ATR and DNA-PKcs to sites of DNA damage. *Nature*, 434, 605–611.

Gaiser,O.J. *et al.* (2004) Solution structure, backbone dynamics, and association behavior of the C-terminal BRCT domain from the breast cancer-associated protein BRCA1. *Biochemistry*, 43, 15983–15995.

Glover,J.N. *et al.* (2004) Interactions between BRCT repeats and phosphoproteins: tangled up in two. *Trends Biochem. Sci.*, 29, 579–585.

Kellis,M. *et al.* (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423, 241–254.

Kobayashi,J. *et al.* (2002) NBS1 localizes to gamma-H2AX foci through interaction with the FHA/BRCT domain. *Curr. Biol.*, 12, 1846–1851.

Kobayashi,J. *et al.* (2004) NBS1 and its functional role in the DNA damage response. *DNA Repair (Amst).*, 3, 855–861.

Lowndes,N.F. and Toh,G.W. (2005) DNA repair: the importance of phosphorylating histone H2AX. *Curr. Biol.*, 15, R99–R102.

Luthy,R. *et al.* (1992) Assessment of protein models with three-dimensional profiles. *Nature*, 356, 83–85.

Manke,I.A. *et al.* (2003) BRCT repeats as phosphopeptide-binding modules involved in protein targeting. *Science*, 302, 636–639.

Petrini,J.H. and Stracker,T.H. (2003) The cellular response to DNA double-strand breaks: defining the sensors and mediators. *Trends Cell Biol.*, 13, 458–462.

Pupko,T. *et al.* (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, 18 (Suppl. 1), S71–S77.

Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, 234, 779–815.

Sippl,M.J. (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins*, 17, 355–362.

Soding,J. (2005) Protein homology detection by HMM–HMM comparison [Erratum, (2005), *Bioinformatics*, 21, 2144]. *Bioinformatics*, 21, 951–960.

Soding,J. *et al.* (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.*, 33, W244–W248.

Stucki,M. *et al.* (2005) MDC1 directly binds phosphorylated histone H2AX to regulate cellular responses to DNA double-strand breaks. *Cell*, 123, 1213–1226.

van den Bosch,M. *et al.* (2003) The MRN complex: coordinating and mediating the response to broken chromosomes. *EMBO Rep.*, 4, 844–849.

Van Der Spoel,D. *et al.* (2005) GROMACS: fast, flexible, and free. *J. Comput. Chem.*, 26, 1701–1718.

Vousden,K.H. and Prives,C. (2005) P53 and prognosis: new insights and further complexity. *Cell*, 120, 7–10.

Wallner,B. and Elofsson,A. (2003) Can correct protein models be identified? *Protein Sci.*, 12, 1073–1086.

Williams,B.R. *et al.* (2002) A murine model of Nijmegen breakage syndrome. *Curr. Biol.*, 12, 648–653.

Yu,X. *et al.* (2003) The BRCT domain is a phospho-protein binding domain. *Science*, 302, 639–642.

Zhang,X. *et al.* (1998) Structure of an XRCC1 BRCT domain: a new protein–protein interaction module. *EMBO J.*, 17, 6404–6411.

Zhao,S. *et al.* (2002) Functional analysis of FHA and BRCT domains of NBS1 in chromatin association and DNA damage responses. *Nucleic Acids Res.*, 30, 4815–4822.

VIII.3.Article 2

Structure, Vol. 12, 1551–1562, September, 2004, ©2004 Elsevier Ltd. All rights reserved. DOI 10.1016/j.str.2004.06.014

The Tudor Tandem of 53BP1: A New Structural Motif Involved in DNA and RG-Rich Peptide Binding

Gaëlle Charier,[1] Joël Couprie,[1] Béatrice Alpha-Bazin,[2] Vincent Meyer,[2] Eric Quéméneur,[3] Raphaël Guérois,[3] Isabelle Callebaut,[4] Bernard Gilquin,[1] and Sophie Zinn-Justin,[1,*]

[1]Département d'Ingénierie et d'Etudes des Protéines
[3]Département de Biologie Joliot-Curie
CEA SACLAY
91191 Gif-sur-Yvette
France

[2]Département d'Ingénierie et d'Etudes des Protéines
CEA VALRHO
30207 Bagnols-sur-Ceze

[4]Département de Biologie Structurale
LMCP, CNRS UMR 7590
4 Place Jussieu
75252 Paris Cedex 05
France

Summary

53BP1 is a key transducer of the DNA damage checkpoint signal, which is required for phosphorylation of a subset of ATM substrates and p53 accumulation. After cell irradiation, the 53BP1 N-terminal region is phosphorylated. Its two C-terminal BRCT motifs interact with p53. Its central region is required and sufficient for 53BP1 foci formation at DNA strand breaks and for 53BP1 binding to the kinetochore. It contains an RG-rich segment and interacts with DNA in vitro. Here we show that the major globular domain of the 53BP1 central region adopts a new structural motif composed of two tightly packed Tudor domains and a C-terminal α helix. A unique surface essentially located on the first Tudor domain is involved in the binding to 53BP1 RG-rich sequence and to DNA, suggesting that the Tudor tandem can act as an adaptor mediating intramolecular as well as intermolecular protein-protein interactions and protein-nucleic acid associations.

Introduction

The DNA double-strand break (DSB) is one of the most serious damages for cells. Such an event may result in rearrangement or loss of genetic information, and lead to cell death or carcinogenesis. DSBs arise during normal endogenous processes of cells (DNA replication, meiosis, V(D)J recombination), but can also be induced by ionizing radiations (IR). In general, cells do not enter S or M phase before the DNA lesions are properly repaired due to the action of the DNA damage checkpoint (Hartwell and Weinert, 1989). The sensitivity of cancer cells to DNA-damaging agents is explained by the fact that cancer cells have often lost some aspects of their checkpoint functions, thus acquiring a higher rate of genomic evolution and a growth advantage (Hartwell and Kastan, 1994).

The mammalian protein p53 Binding Protein 1 (53BP1) was originally identified in a yeast two-hybrid screen as a protein that interacts with p53 DNA binding domain through its two C-terminal BRCT motifs (Iwabuchi et al., 1994; Joo et al., 2002; Derbyshire et al., 2002). BRCT domains are 100–150 residue motifs found in a large number of proteins involved in the cellular responses to DNA damages (Bork et al., 1997; Callebaut and Mornon, 1997a; Clapperton et al., 2004; Williams et al., 2004). Consistently, upon exposure to IR, 53BP1 was shown to rapidly form foci at the sites of DSBs and to be phosphorylated via ATM, a central signaling kinase of the response to DSBs (Schultz et al., 2000; Rappold et al., 2001; Anderson et al., 2001). Further experiments using small-interfering RNA or gene targeting to knockdown 53BP1 expression have shown that 53BP1 is required for the accumulation of p53, for the intra-S-phase and G_2-M checkpoints, and for the phosphorylation of a subset of ATM substrates such as Chk2, BRCA1, and SMC1 in response to IR damage (Wang et al., 2002). These results indicate that 53BP1 is a central mediator of the DNA damage checkpoint.

Radiation-induced phosphorylation of 53BP1 N terminus by ATM kinase is not essential for 53BP1 foci formation. However, the region 1052–1639 of 53BP1, comprised between the N-terminal phosphorylated region and the BRCT domains, is required and sufficient for the recruitment of 53BP1 to DNA strand breaks. In vitro pull-down assays revealed that an overlapping region, comprising residues 956–1354, binds to phosphorylated but not unphosphorylated histone H2AX (Ward et al., 2003). Moreover, several experiments suggested a direct interaction between 53BP1 and phosphorylated H2AX in vivo. First, after irradiation 53BP1 colocalizes with phosphorylated H2AX in megabase regions surrounding the sites of DNA breaks (Schultz et al., 2000; Rappold et al., 2001). Second, phosphorylation of H2AX at serine 140 is critical for efficient 53BP1 foci formation (Ward et al., 2003). Third, H2AX-deficient cells lack normal 53BP1 foci formation and like 53BP1-deficient cells manifest a G_2-M checkpoint defect after exposure to low doses of IR (Fernandez-Capetillo et al., 2002). Thus, recruitment of the central region of 53BP1 to phosphorylated H2AX foci seems to be a crucial step for the initial activation of 53BP1.

53BP1 directly interacts with DNA in vitro. The region 1052–1709 of 53BP1, largely overlapping the region involved in foci formation, binds to a double-strand 10 bp oligonucleotide in blot overlay assays. Furthermore, electrophoresis mobility shift assay experiments suggested that 53BP1 possesses at least two DNA binding domains; both regions 1319–1480 and 1480–1616 bind to linear double- and single-strand DNA substrates. The region 1480–1616 also promotes DNA end joining by the

*Correspondence: szinn@cea.fr

Structure
1552

Table 1. Structural Statistics for the Mouse 53BP1 (1480–1591) Fragment

Number of violations	
nOe distance restraints > 0.5 Å	0
Dihedral restraints > 10°	0
Experimental restraints	
Distance restraints (Å)	2397 (rmsd: 0.04 ± 0.001)
Unambiguous	2081
Ambiguous	256
Hydrogen bonds	60
Dihedral restraints (°)	206 (rmsd: 1.2 ± 0.06)
Rms deviation from idealized covalent geometry	
Bonds (Å)	0.016 ± 0.0003
Angles (°)	3.6 ± 0.04
Impropers (°)	3.0 ± 0.2
Energy (kcal/mol)[a]	
van der Waals	171 ± 11
Electrostatic	−408 ± 22
Ramachandran plot (%)	
Most favored regions	79.8
Additionally allowed regions	18.4
Coordinate precision (residues 7–129)	
Backbone atoms	0.82 ± 0.2
Heavy atoms	1.3 ± 0.16

[a]The van der Waals energy is calculated with a Lennard–Jones potential. The electrostatic energy is calculated with no net charge on side chain atoms and a distance-gated dielectric constant. CHARMM22 parameters were used.

DNA ligase IV/Xrcc4 complex, which is involved in the nonhomologous end joining (NHEJ) pathway of DSB repair in mammalian cells. Thus, the central region of 53BP1 might directly participate to the repair of DNA DSBs (Iwabuchi et al., 2003).

Finally, 53BP1 localizes to kinetochores and is hyperphosphorylated during mitosis under conditions where the spindle checkpoint is activated. The minimal 53BP1 kinetochore binding domain resides again in a region located between the N terminus phosphorylated in an ATM-dependent manner and the BRCT domains. It corresponds to residues 1220–1601 in mouse 53BP1 and 1235–1616 in human 53BP1. Thus, the central region of 53BP1 is involved both in the DNA damage response and in the signaling at the kinetochore during mitosis (Jullien et al., 2002).

We have analyzed the sequence of 53BP1 between residues 956 and 1709. It contains several low complexity regions. However, a segment showing globular domain characteristics (about one-third of hydrophobic residues) is found between residues 1480 and 1616. This segment is predicted to contain a Tudor domain, a conserved motif of 50 amino acids found in several RNA-associated proteins (Ponting, 1997; Maurer-Stroh et al., 2003; Callebaut and Mornon, 1997b). A Tudor domain was also found in the SMN protein, a protein linked to spinal muscular atrophy. Its 3D structure was solved by NMR (Selenko et al., 2001) and by X-ray crystallography (Sprangers et al., 2003). The Tudor domain of SMN binds to symmetrically dimethylated arginines (sDMA) of RG-rich sequences found in Sm protein C-terminal tails (Selenko et al., 2001). Interestingly, an RG-rich motif of 8 residues (RGRGRRGR) is highly conserved within the three known 53BP1 sequences (human, mouse, *Xenopus*). This motif is located 80 residues before the predicted Tudor domain in mouse 53BP1.

To gain molecular insight into the functional role of the region of 53BP1 comprised between the N-terminal phosphorylated region and the BRCT domains, we have solved the three-dimensional structure of the segment 1463–1617 of mouse 53BP1 (53BP1^{TT}) using NMR. Here we show that this segment, which is 99% identical to region 1478–1632 of human 53BP1, surprisingly comprises not only one but two Tudor domains. The first Tudor domain corresponds to the predicted one and presents a cavity analogous to the sDMA binding region of SMN. At this point, NMR measurements were used to characterize the interaction of 53BP1^{TT} with an RGRGRRGR peptide symmetrically dimethylated or nonmethylated on the arginine residues. The DNA binding region of 53BP1^{TT} was also mapped by NMR. These different interactions were analyzed, in order to understand their specificity and their potential mode of regulation. Finally, searching for similar functional sites in the other detected tandem Tudor sequences was carried out in order to understand the functional role of this new structural family.

Results

53BP1^{TT} Folds into Three Structural Motifs

The solution structure of 53BP1^{TT} was determined by heteronuclear double and triple resonance NMR spectroscopy (Table 1). Coordinates and NMR restraints were deposited at the Protein Data Bank (entry code: 1SSF, http://www.rcsb.org/pdb/). In the following, the supplementary glycine resulting from the biomolecular construction is numbered 0 and the last residue 155. 53BP1^{TT} is folded between residues 8 and 129. Indeed, positive ^{1}H-^{15}N nOe were measured between residues 8 and 129 (data not shown), and the backbone root-mean-square deviation (rmsd) calculated on this fragment with respect to the mean coordinates yields 0.85 Å. The 3D structure of region 8–129 is constituted of three structural motifs (Figures 1A and 1B). The first and second motifs correspond to residues 12–57 and 64–112, respectively. They both adopt a β-barrel-like fold and are connected by a 6 residue linker. The third motif essen-

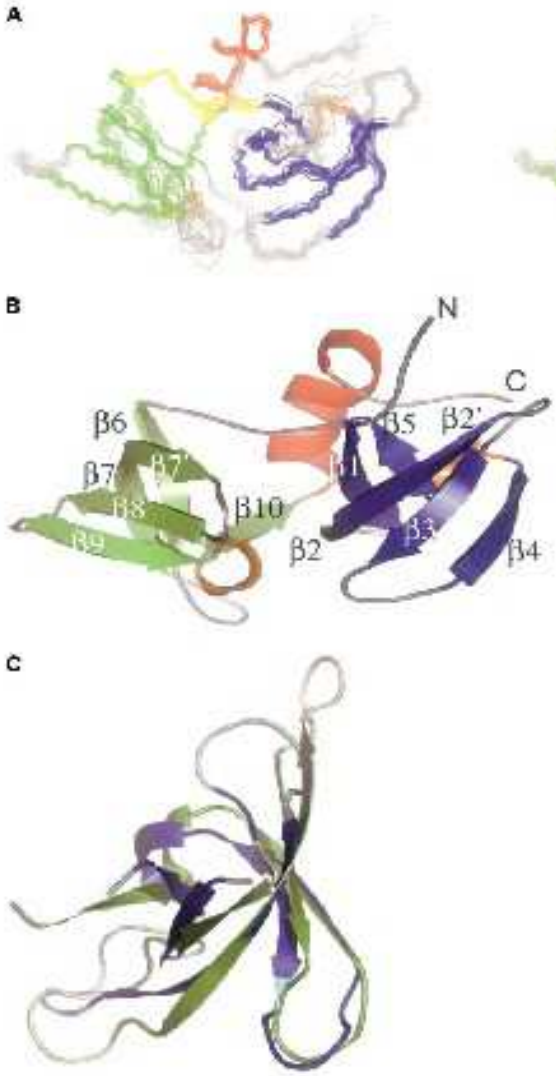


Figure 1. Structure of the Mouse 53BP1 (1483-1617) Region

(A) Stereoview of the backbone atoms (N, C_α, C') for residues 7–129 (corresponding to residues 1469–1591 of mouse 53BP1) of a set of ten superimposed structures. Secondary structures are colored in blue for β strands of Tudor motif 1, green for β strands of Tudor motif 2, orange for 3_{10} helices of motifs 1 and 2, and red for the C-terminal helix.

(B) Ribbon representation of residues 7–129 colored as in (A).

(C) Superimposition of Tudor motifs 1 and 2 on backbone atoms (N, C_α, C'). Dark blue and green correspond to fitted residues (12–17 on 65–70, 24–31 on 76–83, 37–41 on 92–96, and 43–57 on 97–111). Light blue and green correspond to the other residues.

tially corresponds to an α helix located between residues 113 and 123. The three motifs are mostly rigid on a picosecond to nanosecond timescale (^1H-^{15}N nOe > 0.6 for 87% of the residues). Only loop β6β7, in motif 2, and segment 124–129, following the α helix, are clearly more flexible (^1H-^{15}N nOe < 0.4).

Analysis of the NMR spectra revealed that 13 residues from $53BP1^{TT}$ display two sets of resonances with unequal intensities. These residues are spread all over the sequence. In the 3D structure, they are clustered in and around the C-terminal α-helix, close to the *trans* Pro127. This suggests that the minor conformation of $53BP1^{TT}$ corresponds to a *cis* conformer of Pro127. Our structural analysis focuses on the major form of the $53BP1^{TT}$ fragment.

Motifs 1 and 2 Adopt Similar β-Barrel-like Structures

As shown in Figure 1C, motifs 1 and 2 are mostly superimposable, except for two structurally variable loops located between the first and the second strands and between the second and the third strands of each motif.

(two glycines, one alanine, one valine), aliphatic hydrophobic residues (one leucine), or aromatic residues (one phenylalanine, two tyrosines). Val14, Ala16, and Gly27 in motif 1 and Val67, Ala69, and Gly79 in motif 2 are clustered in the two first β strands of their respective motifs; they are directed toward the β-barrel interior and their small size is probably essential for the packing of the two β sheets of their barrel. Gly45 in motif 1 and Gly99 in motif 2 are important for the formation of the β-turns between the third and the fourth β strands. The role of the conserved hydrophobic residues (Tyr23, Phe24, Tyr38, and Leu57 in motif 1 and Tyr75, Phe76, Tyr93, and Leu111 in motif 2) is less clear. All these residues except Tyr75 are buried, and they are found in equivalent positions in the two motifs. However, because the two motifs are juxtaposed in a similar orientation in the 3D structure (Figure 1B), the conserved residues are located differently relative to the interface. Thus, in between motifs 1 and 2, Tyr23 interacts with Phe76 and Phe24 with Leu111, in between the N- and C termini, Tyr38 interacts with Leu57, and finally Tyr93 interacts with the linker.

Motifs 1 and 2 are tightly packed: the second strand of motif 1 and the fifth strand of motif 2 form an antiparallel β sheet through residues 23–25 and 110–112, respectively. Interaction of motif 1 with motif 2 buries a surface of 1540 Å², which is typical for protein-protein complexes (Janin, 1995). Furthermore, the C-terminal α helix participates to the packing of 53BP1^{TT} by interacting with the first, second, and fifth strands of motif 1 and the first and fifth strands of motif 2. Interaction of this helix with motifs 1 and 2 buries 1225 Å². Thus, the total buried area corresponding to the interaction between the three structural motifs is particularly large (2765 Å²), suggesting that a coupling might exist between association and folding of the individual motifs.

Motifs 1 and 2 Correspond to a New Tudor Tandem Fold

A DALI search using motif 1 gives NusG (PDB code: 1M1G) and SMN (PDB code: 1G5V) as the two closest matches. These structures are also found in second and fourth positions when motif 2 is proposed to DALI. Clearly, the β sheet arrangement of motifs 1 and 2 is similar to that found in the Tudor fold of SMN (Selenko et al., 2001) and the Tudor-like fold of NusG (Steiner et al., 2002). If the structures of motifs 1 and 2 are superimposed onto the structure of the SMN Tudor domain, the backbone rmsd calculated on the five β strands yields 2.0 and 1.5 Å, respectively (Figure 2B). All the small and medium size amino acids conserved between motifs 1 and 2 of 53BP1 are also small and medium size amino acids in the SMN Tudor domain (Figure 2A). However, the aromatic residues conserved between motifs 1 and 2 are not found in the SMN Tudor domain, stressing that these residues are probably not crucial for the Tudor fold.

Motif 2 was not predicted as a Tudor domain by SMART (Letunic et al., 2004). The sequence alignment of Figure 2A shows which of the 16 best-conserved positions characterizing a Tudor domain are conserved in SMN and 53BP1 motifs 1 and 2. Clearly, if motif 1 presents most of the markers of a Tudor domain, motif 2 only presents two-thirds of them, and those essentially correspond to buried small or hydrophobic residues. In particular, the solvent-exposed aromatic patch found in SMN and motif 1 is absent in motif 2.

The DALI search using 53BP1 Tudor motifs further revealed a structural analogy with PWWP and MBT domains, evolutionary related to the Tudor domains within the "Royal family" described by Ponting and coworkers (Maurer-Stroh et al., 2003). Finally, the DALI search pointed out the structural analogy between the two first 53BP1 motifs and other SH3-like barrels of various biological functions.

53BP1^{TT} Interacts with the 53BP1 Arg-Gly-Rich Sequence

Alignment of the three known 53BP1 sequences shows that a RGRGRRGR stretch is highly conserved 80 amino acids upstream from the Tudor tandem (Figure 3A). Mouse 53BP1 is methylated in vivo probably through this RG-rich stretch (Y. Adachi, personal communication). To investigate the potential interaction of nonmethylated or symmetrically dimethylated RGRGRRGR peptides with 53BP1^{TT}, we performed NMR titrations. We added each peptide to the ^{15}N-labeled 53BP1^{TT} sample and followed the chemical shift perturbations of 53BP1^{TT} residues by recording ^{1}H-^{15}N HSQC experiments. Figure 3B presents an overlay of the ^{1}H-^{15}N HSQC of 53BP1^{TT} free and saturated with the nonmethylated RG-rich peptide. A similar overlay was obtained after saturation with the methylated peptide (data not shown).

Clearly, 53BP1^{TT} binds to the RGRGRRGR peptides, whether the arginines are nonmethylated or symmetrically dimethylated. The exchange rate between free and bound protein conformations is fast, suggesting that the affinity between 53BP1^{TT} and the peptides is relatively low. Fitting the variation of weighted chemical shift displacements against the peptide concentration yielded a K_d value of 3.9 ± 2.2 mM and 5.8 ± 2.7 mM for the methylated and the nonmethylated peptides, respectively. The estimation of the affinity for the nonmethylated peptide was confirmed by fluorescence measurements (data not show).

Chemical shift changes due to peptide addition are shown in Supplemental Figures S1B and S1C. Residues significantly involved in the interaction with the nonmethylated and the methylated peptides are described in Figures 3C and 3D, respectively. The two mapped interaction surfaces are highly similar. The central peptide binding region (defined by weighted chemical shift displacements higher than 0.05 ppm) contains Trp18, Ser19, Asn21, Gly22, Tyr23 (in loop β1β2), Asp44, Tyr46 (in loop β3β4), Glu47, Cys48 (in strand β4), and Ala69 (in strand β6). Thus, both peptides bind in a cavity essentially composed of three aromatic residues and two negatively charged residues, located between loops β1β2, β3β4 and strand β4 in motif 1. Other residues surrounding this cavity also show substantial chemical perturbations (comprised between 0.025 and 0.05 ppm; Figures 3C and 3D). In particular, several residues of motif 2, located in loop β6β7 and strand β10, are affected by the addition of RG-rich peptides, suggesting that both motifs 1 and 2 are involved in peptide binding.

Structure
1608

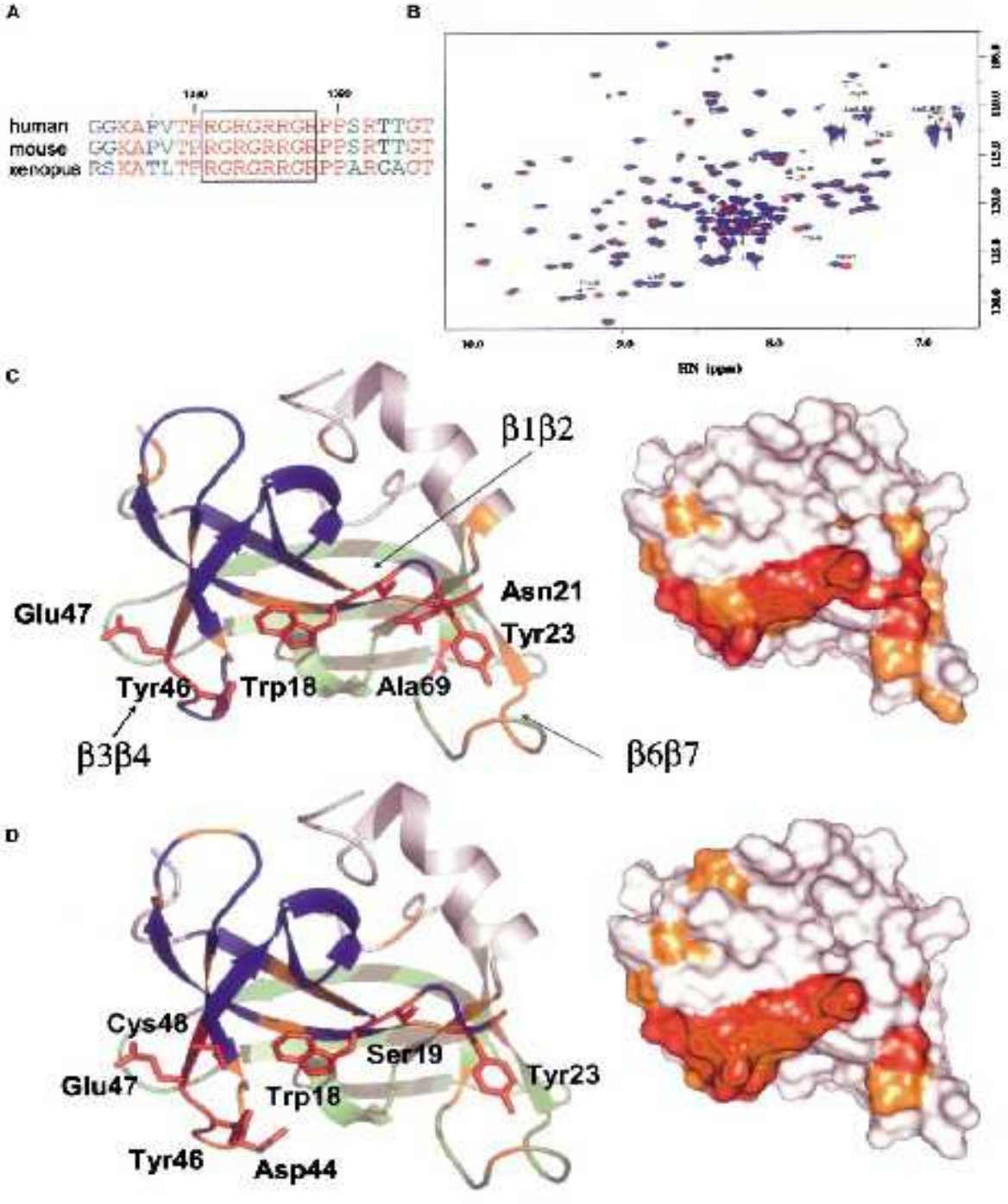


Figure 3. Interaction of 53BP1^{TT} with Two RG-Rich Peptides

(A) Sequence alignment of human, mouse, and *Xenopus* 53BP1 in the RG-rich region (mouse numbering).

(B) Overlay of the ^{1}H-^{15}N HSQC spectra obtained for protein free at 0.6 mM (red) and for protein saturated with the nonmethylated RG-rich peptide (blue).

(C) Interaction with the nonmethylated RG-rich peptide. Left: ribbon representation, with Tudor 1 in blue and Tudor 2 in green. The side chains of residues whose NH-group chemical shift perturbation ($|\Delta\delta(^{1}H)| + 0.1 \times |\Delta\delta(^{15}N)|$) is higher than 0.05 ppm (2 standard deviations) are colored in red and labeled. The backbone is colored in orange if the corresponding chemical shift perturbation is higher than 0.025 ppm. Right: the interaction surface is shown in the same orientation, with the same colors for perturbed amino acids.

(D) Interaction with the symmetrically dimethylated RG-rich peptide. Color codes and orientations are the same as in (C).

In order to characterize the specificity of these interactions, we tested the binding of 53BP1^{TT} to L-arginine alone and to a KG-rich peptide where all the arginines of the initial peptide were substituted by lysines. No substantial chemical shift (>0.02 ppm) variation of 53BP1^{TT} residues was observed after addition of a 50- and 54-fold molar excess of L-arginine and KG-rich peptide, respectively (data not shown). Furthermore, in order to evaluate the potential role of the residues flanking the RG-rich segment in the binding of 53BP1 RG-rich region to 53BP1^{TT}, a 24 residue peptide comprising the RG-rich sequence and relatively well conserved in the three 53BP1 sequences (GGKAPVTPRGRGRRGRPPSR TTGT, nonmethylated, cf. Figure 3A) was added to an NMR sample of 53BP1^{TT}. The longer RG-rich peptide also binds to 53BP1^{TT} with an affinity in the millimolar range: the observed dissociation constant yields 4.7 ± 2.0 mM; the corresponding binding region (weighted chemical shift displacements higher than 0.05 ppm) includes Ala16 (in β1), Trp18, Ser19, Gly22, Tyr23 (in loop β1β2), Phe24, Ser26 (in β2), Phe42 (in β3), Gly45, Tyr46 (in β3β4), Glu47, Cys48 (in β4), Asp59 (in the linker), and Ser71 and Asp73 (in β6β7) (Supplemental Figure S1D).

53BP1^{TT} Is Also Involved in DNA Binding

Doherty and coworkers recently showed that the human 53BP1 (1052-1709) fragment binds to a 10 bp oligonucleotide (5′-AACTCGAGTT-3′) (Iwabuchi et al., 2003). This fragment overlaps the mouse 53BP1 (1463-1617) fragment. The binding of 53BP1^{TT} to this 10 bp oligonucleotide was investigated using NMR. As shown on the ^{1}H-^{15}N HSQC spectra overlay of Figure 4A, addition of DNA induces substantial chemical shift displacements of a set of NH signals, indicating that the 53BP1^{TT} is indeed able to bind to DNA. The dissociation constant estimated for this interaction yields 0.46 ± 0.03 mM.

Figure 4B shows the residues whose NMR signals are strongly perturbed after addition of a 3-fold molar excess of DNA. All chemical shift variations are shown in Supplemental Figure S1A. The central DNA binding region (defined by weighted chemical shift displacements higher than 0.3 ppm) comprises Trp18, Asn21, Gly22, Tyr23, Asp44, and Glu47. Thus DNA binds to 53BP1^{TT} through a surface formed by loops β1β2 and β3β4 in motif 1. Other residues surrounding this central region also show substantial chemical shift displacements (comprised between 0.05 and 0.30 ppm). A larger DNA binding surface can thus be defined involving both Tudor motifs (loops β1β2, β3β4, and β6β7 and strands β4, β6, and β10). Figure 4B shows this large DNA binding surface. Surprisingly, it is similar to the region involved in the binding to RG-rich peptides.

Discussion

The RG-Rich Peptide Binding Site of 53BP1^{TT} Is Essentially Composed of Variable Loops of an SH3-like Fold and Is Largely Superimposable to the Functional Site of SMN

Tudor domains belong to the SH3-like superfold family. SH3-like domains are often part of modular proteins showing a high functional versatility from signal transduction to nucleic acid binding. Their topology is characterized by a five β strand motif and four connecting loops. Large variations in sequence, length, and flexibility of the three first loops are responsible for the functional specificity of the SH3-like domains. In contrast, the fourth loop, which comprises a short 3_{10} helix, is structurally very conserved and plays a key role in preserving the SH3 fold (Dalgarno et al., 1997). In the case of SH3 domains of protein kinases, loop β1β2 is particularly large and is involved, with loop β2β3, in proline-rich peptide binding. In the case of 53BP1 motif 1, loop β1β2 is also involved in the interaction with the targeted RG-rich peptides, but together with the contiguous loop β3β4. In both cases, an aromatic patch is involved in the binding, but this patch is close to loop β2β3 in SH3 domains and close to loop β3β4 in 53BP1.

In contrast, the RG-rich peptide binding site of 53BP1 motif 1 is similar to the binding site of the SMN Tudor domain to Sm protein RG-rich tails. Indeed, using NMR titrations, Sattler and coworkers showed that the amide groups strongly involved in the SMN/poly RG interaction belong to a cluster of conserved aromatic residues: Trp102 (in β1β2), Tyr109 (in β2), Tyr127 (in β3), and Tyr130 (in β3β4). A negatively charged amino acid, Glu134, is also important since the binding to Sm proteins is abolished when this amino acid is mutated to a lysine (Selenko et al., 2001). In the alignment of 53BP1 motif 1 with SMN (Figure 2A), these aromatic residues correspond to Trp18, Tyr25, Phe42, and Asp44, respectively. Glu134 in SMN is aligned with Cys48 in 53BP1. Three and four of these five 53BP1 residues are perturbed by the addition of the unmethylated and the methylated RG-rich peptides, respectively. Tyr25 of 53BP1 is not involved in the binding, but the close Tyr23 is affected by addition of the peptides. Thus, the RG-rich peptide binding site of 53BP1 is mostly superimposable to the SMN functional site, but it is larger because located both in motifs 1 and 2 and involving the more distant Tyr23 (Figure 2C).

Arginine Methylation of the RG-Rich Peptide or Addition of 8 Residues from the 53BP1 Sequence on Both Sides Do Not Influence the Binding to 53BP1^{TT}

SMN and 53BP1 use similar aromatic pockets to bind with a relatively low affinity to nonmethylated and arginine symmetrically dimethylated RG-rich peptides. Sattler and coworkers showed that arginine methylation increases the affinity of the SMN Tudor domain for RG repeats contained in the C terminus of Sm proteins (Sprangers et al., 2003). However, SMN efficiently binds some other RG-rich containing substrates, such as fibrillarin, nucleolin GAR1, Sm core proteins hnRNP Q, R, and U, RNA helicase A, and p80 coilin (Young et al., 2003). In their paper, Lorson and coworkers reported that SMN Tudor domain binds to the nonmethylated RG-rich peptide from the Ewing's sarcoma protein (EWS) with a dissociation constant of 3 mM and to the symmetrically dimethylated peptide with a dissociation constant of 5 mM (Young et al., 2003). We report similar values for the binding of 53BP1^{TT} to nonmethylated and symmetrically dimethylated RG-rich peptides (affinity

Structure
1058

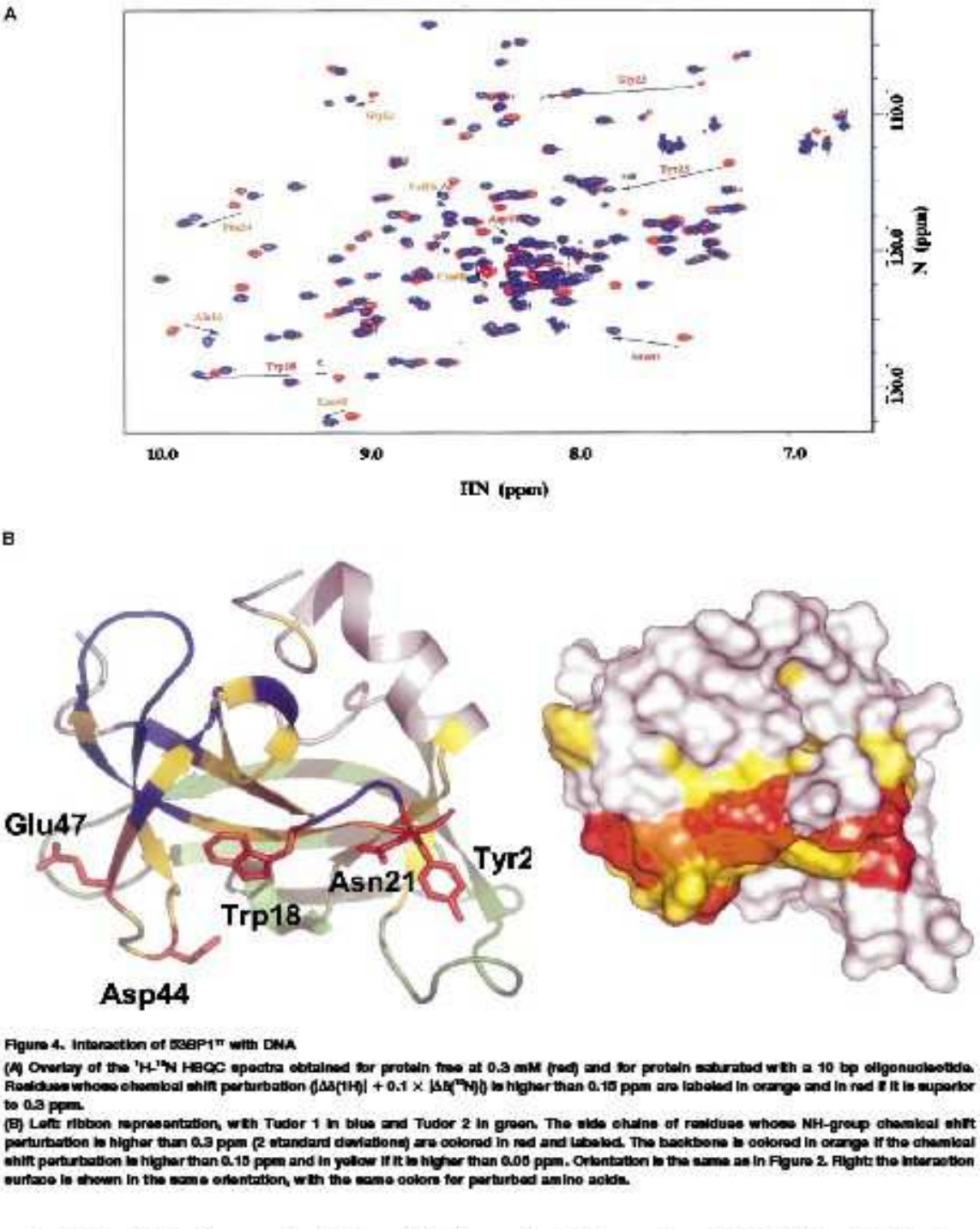


Figure 4. Interaction of 53BP1^{TT} with DNA

(A) Overlay of the ^{1}H-^{15}N HSQC spectra obtained for protein free at 0.3 mM (red) and for protein saturated with a 10 bp oligonucleotide. Residues whose chemical shift perturbation ($|\Delta\delta(^{1}H)| + 0.1 \times |\Delta\delta(^{15}N)|$) is higher than 0.15 ppm are labeled in orange and in red if it is superior to 0.3 ppm.

(B) Left: ribbon representation, with Tudor 1 in blue and Tudor 2 in green. The side chains of residues whose NH-group chemical shift perturbation is higher than 0.3 ppm (2 standard deviations) are colored in red and labeled. The backbone is colored in orange if the chemical shift perturbation is higher than 0.15 ppm and in yellow if it is higher than 0.05 ppm. Orientation is the same as in Figure 2. Right: the interaction surface is shown in the same orientation, with the same colors for perturbed amino acids.

values: 3.9 and 5.8 mM, respectively). Thus, in both cases, arginine symmetrical dimethylation is not critical for the interaction. Nevertheless, the presence of arginine in the peptide is critical, as shown by the absence of measurable affinity of the RG-rich peptide for 53BP1^{TT}, and the peptide sequence is important because arginine alone is also not capable to bind to 53BP1^{TT} with a measurable affinity.

We have also tested the affinity of 53BP1^{TT} for a 24 residue peptide of 53BP1 containing the already tested 8 residue RG-rich core, and relatively well conserved in the three 53BP1 sequences. This peptide binds to

53BP1^{TT} on the same surface and with a similar affinity (K_d ~4.7 ± 2.0 mM), as compared to the initial peptide (K_d ~5.8 ± 2.7 mM). This suggests that the flanking residues are not involved in the observed interaction.

Both Tudor Motifs of 53BP1 Are Necessary for DNA Binding

Tandem-arranged protein-interaction modules with restrained orientations have already been observed, in which each domain alone retains its structural integrity and peptide ligand binding activity (Hatada et al., 1995; Hof et al., 1998; Ottinger et al., 1998; Jacobson et al., 2000). By contrast, in the case of mouse 53BP1^{TT}, the DNA and peptides binding sites are centered on an aromatic cavity located in motif 1 (Trp18, Tyr23, Tyr46). These aromatic residues are not conserved in motif 2, justifying the absence of an equivalent binding site in motif 2.

Binding of DNA to 53BP1 was extensively studied by Doherty and coworkers (Iwabuchi et al. 2003). They showed that the region 1480–1616 of human 53BP1 (corresponding to our mouse 53BP1^{TT}) binds to single-strand and double-strand DNA in vitro. On the contrary, they found no interaction between these DNA substrates and the fragments 1480–1540 and 1540–1616 that correspond to isolated Tudor domains 1 and 2 respectively. Moreover, they reported that both Tudor motifs are needed to form foci after X irradiation in cells. As shown by our NMR titration, several residues of motif 2 close to the aromatic cavity of motif 1 are also affected by the addition of DNA (Figure 4B), suggesting that motif 2 is also involved in DNA binding. Our results are thus in agreement with Doherty and coworkers' biological data suggesting that both Tudor domains are needed for interaction with DNA.

Very recently, Zhang and coworkers have reported the three-dimensional structure of tandem PDZ in the rat GRIP1 (Feng et al., 2003). The tandem adopts a compact and stable structure, while the first repeat alone is less stable and its 3D structure is distorted. Mutual stabilization of the two repeats allows ligand binding by the second repeat. Similarly, the tandem BRCT of BRCA1 behave as a single stable fragment in limited proteolysis and X-ray crystallographic studies (Williams et al., 2001), and both repeats are needed for phospho-specific peptide interaction (Manke et al., 2003). Our Tudor tandem structure might be another example where tandem-arranged modules represent functional supramodules with distinct structures and biological functions with respect to individual domains.

A Similar Region of 53BP1^{TT} for DNA and RG-Rich Peptide Binding

Surprisingly, the DNA binding surface is similar to the region involved in the binding to RG-rich peptides. However, DNA is mostly negatively charged and the peptide is largely positively charged. Two remarks can be done to understand this result.

First, among the residues involved in DNA binding, no positively charged residue (Arg, Lys) is found, as is usually the case in protein-DNA interfaces. We compared our results with the analysis of 129 protein-DNA complex structures made by Thornton and coworkers; only three residues (Asn21, Asp44, and Glu47) correspond to typical amino acids forming hydrogen bonds at protein-DNA interfaces (Luscombe et al. 2001). Furthermore, in the same study, the authors noticed that cysteines have a high propensity to contact DNA backbones; in our case, the NMR signal of Cys48 is significantly perturbed by the binding to DNA. Finally, two phenylalanines (Phe 24, Phe42) are also perturbed by the addition of DNA; this amino acid has a high affinity for many DNA base types, which can be explained by its ability to produce extensive ring-stacking interactions (Luscombe et al. 2001). These observations suggest that the 53BP1^{TT}-DNA interface essentially involves hydrogen bonding to nonpositively charged residues and base stacking to aromatic residues.

Second, the RG-rich peptide binding surface is mainly composed of hydrophobic residues (Trp18, Tyr23, Phe24, Phe42, Tyr46); these are able to develop cation-π interactions with the arginines of the peptide (Zacharias and Dougherty, 2002). The Tudor domain of SMN possesses a similar aromatic cavity. However, in the case of SMN, an arginine probably interacts with the negatively charged residue Glu134 of the cavity. In 53BP1^{TT}, no negatively charged side chain is found in the binding site.

To sum up, both DNA and RG-rich peptides interact with 53BP1^{TT} mainly via hydrophobic contacts, cation-π interactions, ring stacking, and hydrogen bonding. No positively charged residue of 53BP1^{TT}is clearly involved in DNA binding and no negatively charged residue is observed in the RG-rich peptide binding site. This may explain why both interactions involve the same region of 53BP1^{TT}.

Tudor Tandem Is Associated to Gene Transcriptional Regulation, Chromatin Remodeling, and DNA Repair

The three-dimensional structure described in this paper is the first solved structure of a tandem of Tudor domains. Many Tudor proteins contain multiple Tudor repeats, and Tudor domains are often arranged in closely linked pairs (Letunic et al., 2004). For example, the human ESET protein possesses two Tudor domains connected by a 38 residue linker, while the human GASC1 protein possesses two such domains connected by an 11 residue linker. Alignment of the sequences of these Tudor tandem domains shows that 28 amino acids of 53BP1^{TT} are conserved in more than 70% of the sequences (Figure 5). Twelve residues are buried and constitute the hydrophobic core of motif 1. Eight residues play a similar structural role in motif 2. Seven residues are localized at the interface between motif 1 and motif 2: Gly11, Arg13, Gly22, Phe24, Tyr25, Asp43, and Arg106. Such a distribution of conserved residues suggests that the role of these residues is mainly to stabilize the Tudor tandem fold.

The Tudor tandem is mostly associated with SET, MBD, PHD, and Zinc Finger C2H2 or BRCT domains, which suggests that their biological function is always linked to gene transcriptional regulation, chromatin remodeling, and DNA repair (Letunic et al., 2004). However, most residues of 53BP1^{TT} involved in RG-rich pep-

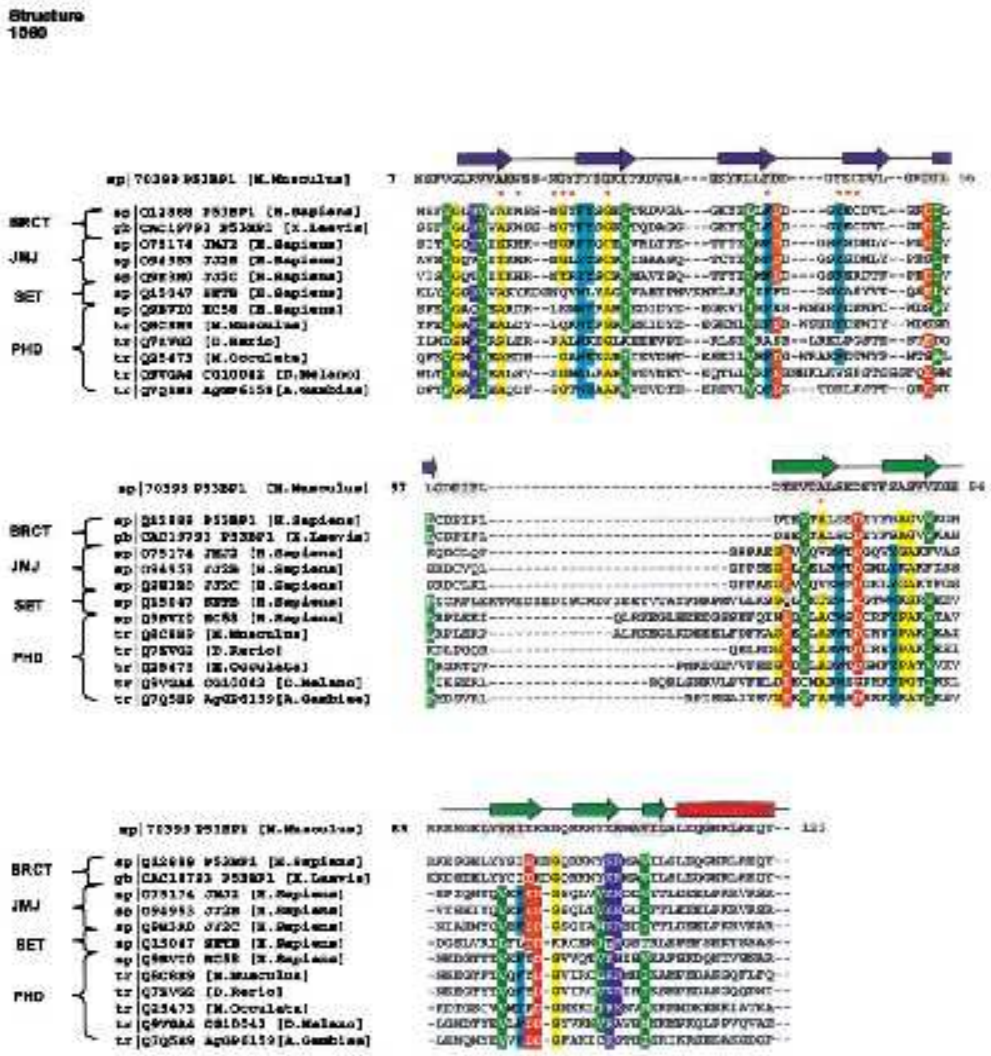


Figure 5. Sequence Alignment of Tudor Tandem Containing Proteins

Conserved residues are colored (dark green for hydrophobic, red for negatively charged, blue for positively charged, cyan for aromatic, and yellow for small residues). Secondary structures of 53BP1^{TT} are indicated (arrows for β strand and rectangle for α helix); mouse 53BP1 residues colored in gray have a relative accessible surface below 20%. Red asterisks show the residues whose NMR signal is affected by the binding to both RG-rich peptides and DNA.

tide or DNA binding are not conserved within other Tudor tandem sequences. Thus, by now, no evidence indicates that Tudor tandems show common target binding properties.

Conclusion

The Tudor domains of SMN and 53BP1 are able to bind to proteins containing RG-rich sequences. Moreover, the ribosomal protein RL24 of *H. marismortui* has a Tudor-like fold and interacts with RNA (Kyrpides et al., 1996; Steiner et al., 2002; Ban et al., 2000). By comparing the RNA binding site of RL24 and the RG-rich peptide binding site of SMN, Wahl and coworkers (Steiner et al., 2002) proposed that the Tudor-like domain of the microbial transcription modulator NusG could bind concomitantly to proteins and nucleic acids via different surfaces. On the opposite, our analysis shows that the Tudor tandem of 53BP1 uses the same surface to bind both peptides and DNA. The presence of this multipartner binding surface suggests that 53BP1^{TT} acts as an adaptor mediating both protein-protein and protein-DNA interactions.

Furthermore, since the tested RG-rich sequence belongs to 53BP1 itself, 53BP1^{TT} may be involved in intramolecular or intermolecular associations. An intramolecular interaction may allow a regulation of the accessibility to other partners of the 53BP1^{TT} functional site. An intermolecular association could entail the accumulation of 53BP1 molecules and, therefore, be a key step in the mechanism of nuclear foci formation observed after cell irradiation. Further studies should be performed to search for other partners of 53BP1^{TT} in the nucleus and to investigate the role of methylation in the interactions between 53BP1^{TT} and RG-rich proteins.

Experimental Procedures

NMR Spectroscopy

53BP1^{TT} was expressed and purified as described previously (Charier et al., 2004). NMR samples were prepared in Tris-HCl 50 mM buffer (pH 7.2) containing 100 mM NaCl in either 99% H_2O/1%

D2O or in 100% D2O. 1 mM EDTA, a protease inhibitor cocktail (SIGMA), 1 mM NaN_3, and 1 mM 3-(trimethylsilyl)[2,2,3,3-$^2H_4$] propionate (TSP) were added. All assignment experiments were performed at 27°C on Bruker DRX-500, DRX-600 equipped with triple-resonance probes according to the previously reported procedure (Charlier et al., 2004). The nOe crosspeak volumes used for structure calculation were measured on four NOESY experiments (a ^{15}N-HSQC-NOESY recorded at 750 MHz at the European Large Scale Facilities in Utrecht, Netherlands, and a ^{13}C-HSQC-NOESY in D_2O, a ^{13}C-HSQC-NOESY in H_2O, and a ^{13}C-HSQC-NOESY in the ^{13}C aromatic region all three recorded on a local 600 MHz spectrometer equipped with a triple resonance TXI cryoprobe). ϕ torsion angle values were deduced from the analysis of the Hn-Hα and the HMQC-J experiments (Vuister and Bax, 1993; Kuboniwa et al., 1994). Hydrogen bound restraints were derived from slowly exchanging amide protons, identified after exchange of H_2O to D_2O followed on ^1H-^{15}N HSQC spectra recorded at different times. All spectra were processed with the programs Xwinnmr (Bruker) or NMRPipe (Delaglio et al., 1995) and analyzed using Felix (Molecular Simulations).

Structure Determination

The solution structure of region 7–129 was solved on the basis of 2337 interproton distances deduced from the nOe data. These distances were estimated from 3365 integrated peak volumes obtained from the four NOESY experiments (935 on the ^{15}N-HSQC-NOESY, 473 on the ^{13}C-HSQC-NOESY in D_2O, 1685 on the ^{13}C-HSQC-NOESY in H_2O, and 270 on the ^{13}C-HSQC-NOESY in the ^{13}C aromatic region). A semiautomated iterative assignment procedure was applied for the assignment and the construction of the 3D structures (Bavarin et al., 2001). A force field adapted to NMR structure calculation (file parallhdg.pro in CNS 1.0 [Brunger et al., 1998]) was used. 206 torsion angles (ϕ or Ψ) values were deduced from the analysis of the Hn-Hα and the HMQC-J experiments and from the backbone ^1H, ^{15}N, and ^{13}C chemical shifts using the program TALOS (Cornilescu et al., 1999). Finally, 30 hydrogen bounds were imposed during the structure calculation. At the last step, 1000 structures were calculated and the 10 best structures were selected and refined with a standard energy function (CHARMM22), including an electrostatic energy term. This term is calculated with no net charge on the side chain atoms and with a distance-gated dielectric constant.

Titrations with Peptides and DNA

NMR titrations were carried out by recording ^1H-^{15}N HSQC experiments at 500 and 600 MHz, using ^{15}N-labeled protein samples at concentrations of 0.3–0.8 mM. The four peptides: RGRGRRGR, R^{a}GR^{a}GR^{a}R^{a}GR^a (R^a = symmetrically dimethylated R), KGKGKKGK, and GGKAPVTPRGRGFRGRPPSRTTGT, from Peptide Specialty Laboratories GmbH (Heidelberg), were added up to 9-, 6-, 54-, and 34-fold molar excess to the protein samples, respectively. The 10 bp oligonucleotide 5′-AACTCGAGTT-3′ (PROLIGO, Paris) was annealed prior to NMR experiments and added up to a 3-fold molar excess to the protein sample. Assuming that the exchange rate is fast and that these interactions show low affinities, we can consider that the concentration of the free ligand is approximately the same as the concentration of the added ligand. Thus dissociation constants were estimated by fitting the titration curves obtained for strongly perturbed residues with Kaleidagraph software, using the approximate equation $y = \Delta\delta_{max} \times x/(K_d + x)$, where y is the weighted chemical shift displacements $|\Delta\delta(^1H)| + 0.1 \times |\Delta\delta(^{15}N)|$ ($\Delta\delta = \delta_{exp} - \delta_0$, δ_{exp} is the observed chemical shift, δ_0 is the initial chemical shift before adding the ligand), x is the ligand concentration, $\Delta\delta_{max}$ is the maximum variation of the weighted chemical shift displacements, and K_d is the estimated dissociation constant.

Bio-Informatics

An initial multiple alignment was built from Tudor tandem sequences found in the SMART database (Letunic et al., 2004) (http://smart.embl-heidelberg.de/smart/). The sequence of 53BP1^{TT} was structurally aligned to this set of sequences. Using HMMER2.3 (Durbin et al., 1998) (http://hmmer.wustl.edu/), a HMM profile was built and used to scan the Non Redundant database (ftp://ftp.ncbi.nih.gov/blast/db/). New sequences were detected (E-value < $1e^{-30}$) and added to the initial alignment. Sequences with high identity were excluded. The final alignment was manually optimized using the secondary structure of 53BP1^{TT}.

Supplemental Data

Histograms representing the absolute value of chemical shift displacements of the whole residues of 53BP1^{TT} after saturation with DNA or different RG-rich peptides are provided as Supplemental Data and can be found at http://www.structure.org/cgi/content/full/12/9/1551/DC1.

Acknowledgments

The 750 MHz spectrum was recorded at the SON NMR Large Scale Facility in Utrecht, which is funded by the "Access to Research Infrastructures" program of the European Union (HPRI-CT-2001-00172). We are grateful to Philippe Bavarin and Flavio Toma who kindly lent us their 600 MHz spectrometer. We thank D. Julien for providing mouse 53BP1 cDNA and G. Stier for TEV protease cDNA. Vincent Meyer is supported by grants from Association Française contre les Myopathies (AFM, grant Decrypton 2003 #9457).

Received: March 31, 2004
Revised: June 16, 2004
Accepted: June 19, 2004
Published: September 7, 2004

References

Anderson, L., Henderson, C., and Adachi, Y. (2001). Phosphorylation and rapid relocalization of 53BP1 to nuclear foci upon DNA damage. Mol. Cell. Biol. *21*, 1719–1729.

Ban, N., Nissen, P., Moore, P.B., and Steitz, T.A. (2000). The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. Science *289*, 905–920.

Bork, P., Hofmann, K., Bucher, P., Neuwald, A.F., Altschul, S.F., and Koonin, E.V. (1997). A superfamily of conserved domains in DNA damage-responsive cell cycle checkpoint proteins. FASEB J. *11*, 68–76.

Brunger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., et al. (1998). Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr. D Biol. Crystallogr. *54*, 905–921.

Callebaut, I., and Mornon, J.P. (1997a). From BRCA1 to RAP1: a widespread BRCT module closely associated with DNA repair. FEBS Lett. *400*, 25–30.

Callebaut, I., and Mornon, J.P. (1997b). The human EBNA-2 coactivator p100: multidomain organization and relationship to the staphylococcal nuclease fold and to the tudor protein involved in Drosophila melanogaster development. Biochem. J. *321*, 125–132.

Charlier, C., Alpha-Bazin, B., Couprie, J., Callebaut, I., Berenguer, F., Quemeneur, E., Gilquin, B., and Zinn-Justin, S. (2004). Letter to the editor: 1H, 13C and 15N resonance assignments of the region 1463–1617 of the mouse p53 Binding Protein 1 (53BP1). J. Biomol. NMR *28*, 303–304.

Clapperton, J.A., Manke, I.A., Lowery, D.M., Ho, T., Haire, L.F., Yaffe, M.B., and Smerdon, S.J. (2004). Structure and mechanism of BRCA1 BRCT domain recognition of phosphorylated BACH1 with implications for cancer. Nat. Struct. Mol. Biol. *11*, 512–518.

Cornilescu, G., Delaglio, F., and Bax, A. (1999). Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J. Biomol. NMR *13*, 289–302.

Dalgarno, D.C., Botfield, M.C., and Rickles, R.J. (1997). SH3 domains and drug design: ligands, structure, and biological function. Biopolymers *43*, 383–400.

Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J., and Bax, A. (1995). NMRPipe: a multidimensional heteronuclear processing system based on UNIX pipes. J. Biomol. NMR *6*, 277–293.

Derbyshire, D.J., Basu, B.P., Serpell, L.C., Joo, W.S., Date, T., Iwabuchi, K., and Doherty, A.J. (2002). Crystal structure of human 53BP1 BRCT domains bound to p53 tumour suppressor. EMBO J. *21*, 3863–3872.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids (Cambridge, UK: Cambridge University Press).

Feng, W., Shi, Y., Li, M., and Zhang, M. (2003). Tandem PDZ repeats in glutamate receptor-interacting proteins have a novel mode of PDZ domain-mediated target binding. Nat. Struct. Biol. *10*, 972–978.

Fernandez-Capetillo, O., Chen, H.T., Celeste, A., Ward, I., Romanienko, P.J., Morales, J.C., Naka, K., Xia, Z., Camerini-Otero, R.D., Motoyama, N., et al. (2002). DNA damage-induced G2-M checkpoint activation by histone H2AX and 53BP1. Nat. Cell Biol. *4*, 993–997.

Hartwell, L.H., and Kastan, M.B. (1994). Cell cycle control and cancer. Science *266*, 1821–1828.

Hartwell, L.H., and Weinert, T.A. (1989). Checkpoints: controls that ensure the order of the cell cycle events. Science *246*, 629–634.

Hatada, M.H., Lu, X., Laird, E.R., Green, J., Morgenstern, J.P., Lou, M., Marr, C.S., Phillips, T.B., Ram, M.K., Theriault, K., et al. (1995). Molecular basis for interaction of the protein tyrosine kinase ZAP-70 with the T-cell receptor. Nature *377*, 32–38.

Hof, P., Pluskey, S., Dhe-Paganon, S., Eck, M.J., and Shoelson, S.E. (1998). Crystal structure of the tyrosine phosphatase SHP-2. Cell *92*, 441–450.

Iwabuchi, K., Bartel, P.L., Li, B., Marraccino, R., and Fields, S. (1994). Two cellular proteins that bind to wild-type but not mutant p53. Proc. Natl. Acad. Sci. USA *91*, 6098–6102.

Iwabuchi, K., Basu, B.P., Kysela, B., Kurihara, T., Shibata, M., Guan, D., Cao, Y., Hamada, T., Imamura, K., Jeggo, P.A., et al. (2003). Potential role for 53BP1 in DNA end-joining repair through direct interaction with DNA. J. Biol. Chem. *278*, 36487–36495.

Jacobson, R.H., Ladurner, A.G., King, D.S., and Tjian, R. (2000). Structure and function of a human TAFII250 double bromodomain module. Science *288*, 1422–1425.

Janin, J. (1995). Elusive affinities. Proteins *21*, 30–39.

Joo, W.S., Jeffrey, P.D., Cantor, S.B., Finnin, M.S., Livingston, D.M., and Pavletich, N.P. (2002). Structure of the 53BP1 BRCT region bound to p53 and its comparison to the Brca1 BRCT structure. Genes Dev. *16*, 583–593.

Jullien, D., Vagnarelli, P., Earnshaw, W.C., and Adachi, Y. (2002). Kinetochore localisation of the DNA damage response component 53BP1 during mitosis. J. Cell Sci. *115*, 71–79.

Kuboniwa, H., Grzesiek, S., Delaglio, F., and Bax, A. (1994). Measurement of HN-H alpha J couplings in calcium-free calmodulin using new 2D and 3D water-flip-back methods. J. Biomol. NMR *4*, 871–878.

Kyrpides, N.C., Woese, C.R., and Ouzounis, C.A. (1996). KOW: a novel motif linking a bacterial transcription factor with ribosomal proteins. Trends Biochem. Sci. *21*, 425–426.

Letunic, I., Copley, R.R., Schmidt, S., Ciccarelli, F.D., Doerks, T., Schultz, J., Ponting, C.P., and Bork, P. (2004). SMART 4.0: towards genomic data integration. Nucleic Acids Res. *32*, D142–D144.

Luscombe, N.M., Laskowski, R.A., and Thornton, J.M. (2001). Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic Acids Res. *29*, 2860–2874.

Manke, I.A., Lowery, D.M., Nguyen, A., and Yaffe, M.B. (2003). BRCT repeats as phosphopeptide-binding modules involved in protein targeting. Science *302*, 636–639.

Maurer-Stroh, S., Dickens, N.J., Hughes-Davies, L., Kouzarides, T., Eisenhaber, F., and Ponting, C.P. (2003). The Tudor domain 'Royal Family': Tudor, plant Agenet, Chromo, PWWP and MBT domains. Trends Biochem. Sci. *28*, 69–74.

Ottinger, E.A., Botfield, M.C., and Shoelson, S.E. (1998). Tandem SH2 domains confer high specificity in tyrosine kinase signaling. J. Biol. Chem. *273*, 729–735.

Ponting, C.P. (1997). Tudor domains in proteins that interact with RNA. Trends Biochem. Sci. *22*, 51–52.

Rappold, I., Iwabuchi, K., Date, T., and Chen, J. (2001). Tumor suppressor p53 binding protein 1 (53BP1) is involved in DNA damage-signaling pathways. J. Cell Biol. *153*, 613–620.

Savarin, P., Zinn-Justin, S., and Gilquin, B. (2001). Variability in automated assignment of NOESY spectra and three-dimensional structure determination: a test case on three small disulfide-bonded proteins. J. Biomol. NMR *19*, 49–62.

Schultz, L.B., Chehab, N.H., Malikzay, A., and Halazonetis, T.D. (2000). p53 binding protein 1 (53BP1) is an early participant in the cellular response to DNA double-strand breaks. J. Cell Biol. *151*, 1381–1390.

Selenko, P., Sprangers, R., Stier, G., Buhler, D., Fischer, U., and Sattler, M. (2001). SMN Tudor domain structure and its interaction with the Sm proteins. Nat. Struct. Biol. *8*, 27–31.

Sprangers, R., Groves, M.R., Sinning, I., and Sattler, M. (2003). High-resolution X-ray and NMR structures of the SMN Tudor domain: conformational variation in the binding site for symmetrically dimethylated arginine residues. J. Mol. Biol. *327*, 507–520.

Steiner, T., Kaiser, J.T., Marinkovic, S., Huber, R., and Wahl, M.C. (2002). Crystal structures of transcription factor NusG in light of its nucleic acid- and protein-binding activities. EMBO J. *21*, 4641–4653.

Vuister, G.W., and Bax, A. (1993). Quantitative J correlation: a new approach for measuring homonuclear three-bond J(HN-Ha) coupling constants in 15N-enriched proteins. J. Am. Chem. Soc. *115*, 7772–7777.

Wang, B., Matsuoka, S., Carpenter, P.B., and Elledge, S.J. (2002). 53BP1, a mediator of the DNA damage checkpoint. Science *298*, 1435–1438.

Ward, I.M., Minn, K., Jorda, K.G., and Chen, J. (2003). Accumulation of checkpoint protein 53BP1 at DNA breaks involves its binding to phosphorylated histone H2AX. J. Biol. Chem. *278*, 19579–19582.

Williams, R.S., Green, R., and Glover, J.N. (2001). Crystal structure of the BRCT repeat region from the breast cancer-associated protein BRCA1. Nat. Struct. Biol. *8*, 838–842.

Williams, R.S., Lee, M.S., Hau, D.D., and Glover, J.N.M. (2004). Structural basis of phosphopeptide recognition by the BRCT domain of BRCA1. Nat. Struct. Mol. Biol. *11*, 519–525.

Young, P.J., Francis, J.W., Lince, D., Coen, K., Androphy, E.J., and Lorson, C.L. (2003). The Ewing's sarcoma protein interacts with the Tudor domain of the survival motor neuron protein. Brain Res. Mol. Brain Res. *119*, 37–49.

Zacharias, N., and Dougherty, D.A. (2002). Cation-pi interactions in ligand recognition and catalysis. Trends Pharmacol. Sci. *23*, 281–287.

Accession Numbers

The NMR restraints and protein coordinates have been deposited in the PDB under entry code 1S8F. The chemical shifts have been deposited at the BMRB under accession number 5578.

VIII.4.Article 3 (en préparation)

Application note

HMM-Kalign: a tool for generating sub-optimal HMM alignments

Emmanuelle Becker*, Aurélie Cotillard, Vincent Meyer, Hocine Madaoui and Raphaël Guérois*

CEA, iBiTecS, URA CNRS 2096, Laboratoire de Biologie Structurale et Radiobiologie, Gif Sur Yvette, F-91191 France.

ABSTRACT

Summary: Recent development of strategies using multiple sequence alignments (MSA) or profiles to detect remote homologies between proteins has led to a significant increase in the number of proteins whose structures can be generated by comparative modelling methods. However, prediction of the optimal alignment between these highly divergent homologous proteins remains a difficult issue. We present a tool based on a generalized Viterbi algorithm that generates optimal and sub-optimal alignments between one sequence and one HMM. The tool is implemented it as a new function within the HMMER package called *hmmkalign*.

Availability: freely available at www.spider.cea.fr.

Contacts: raphael.guerois@cea.fr, emmanuelle.becker@cea.fr.

The present work aims at automatically exploring the alignment space in the neighborhood of the optimal sequence alignment (OSA) in order to find an alignment closer to the structural alignment than the OSA.

The sequence alignment space in the neighborhood of the OSA has been quite extensively explored in the context of pairwise sequence alignments. Waterman proposed an algorithm derived from the standard Sellers algorithm to determine all the pairwise alignments whose scores are within a range ε of the OSA's score. Later and still dealing with pairwise sequence alignments, Saqi and Sternberg proposed a heuristic known as the Iterative Elimination Method, based on the progressive perturbation of the distance matrix. Another method to generate alternative pairwise sequence alignments has been introduced by Zucker .

With the rising of sequence-profile, sequence-HMM, and more recently profile-profile and HMM-HMM alignments, this algorithmic studies were left background. However, although progresses have been made especially for the detection of remote homology, the alignment of sequences sharing less than 25% sequence identity is still problematic in the context of comparative modelling. Based on this observation, two articles came back to the idea of generating alternative alignments and use two heuristics : a parametric approach coupled with Saqi and Sternberg's Iterative Elimination Method, and a genetic algorithm, respectively.

In this work, we explore the possibility of generating alternative alignments in the context of alignments obtained using Hidden Markov Models, such as HMMER or SAM . Instead of heuristics, HMM-Kalign generates the exact neighborhood of the OSA.

The Viterbi algorithm is classically used to align a sequence s_{obs} to a profile HMM and consists in finding the sequence of states that maximizes the emission probability of s_{obs} . To generate alternative alignments in the neighborhood of the OSA, one solution is to use a generalized Viterbi algorithm that precisely determines the k-best sequences of states that maximizes the emission of s_{obs}. This generalization of the Viterbi algorithm has been used in the field of speech recognition and elegant variants have been developed recently that fasten the process . We implemented and included the generalized Viterbi algorithm in the program HMMER.

1 GENERATING SUB-OPTIMAL ALIGNMENTS

To use the *hmmkalign* command, two files are required :

- <*MSA*>, that contains a multiple sequence alignment (derived for instance from the alignment of structural templates);
- <*sequences*>, that contains two sequences in fasta format (i) the sequence to be aligned, (ii) one sequence from the <MSA> file that may be used as a template to further build a model of the first sequence.

To build the HMM, it is possible to use the classical command :

$ *./hmmbuild* <*hmm file* > <*MSA*> (command 1)

although our results show that within highly divergent families, it is more effective to drive explicitly the HMM architecture with respect to the conservation of the secondary structures (details in supplementary data). This is possible via the command :

$ *./hmmbuild --hand* <*hmm file* > <*MSA*> (command 2)

where the <*MSA*> file contains an additional line with symbols '-' and 'x' encoding for the positions of insertions and match states, respectively. After having created the HMM, the command to generate *k* alignments is :

$ *./hmmkalign k* <*hmm file*> <*sequences*>

The OSA classically generated with HMMER corresponds to the alignment with the best score (K=1) (cf. command 1).

Exploration can be targeted to specific regions. For a sequence s_{obs}=s_1...s_T in which only the region s_i...s_j is to be sampled, add a hybrid sequence in the <*MSA*> file, that contains the "anchors" s_1...s_{i-1} and s_{j+1}...s_T and insertions '-' symbols instead of s_i...s_j.

2 TESTING PROCEDURE

We studied 115 alignments from 22 highly divergent protein families, *i.e.* sharing on average less than 25% identity (details in supplementary data). These alignments were extracted from the HOMSTRAD database which contains multiple structural alignments from a large set of families . The following procedure was applied : (1) exclude the test sequence from the multiple structural alignment; (2) build two distincts HMMs with command 1 and command 2, (3) align the excluded sequence to the two HMM with hmmkalign to generate

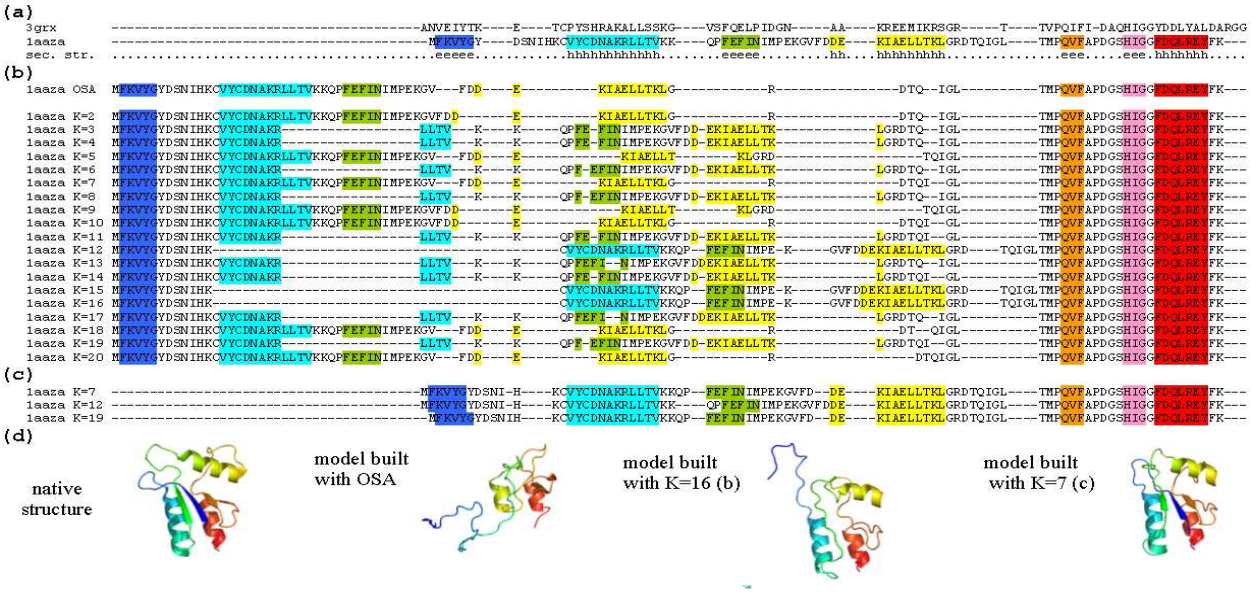


Figure 1 : Oxidized bacteriophage T4 glutaredoxin (1AAZ). The multiple alignments of the sequence with the other members of the thioredoxin family are represented through a projection into a pairwise alignment between 1aaza and 3grx. The amino acids of 1aaza corresponding to a secondary structure are highlighted in color. **(a)** The first two lines present the structural alignment and the secondary structure assigned to 1aaza (HOMSTRAD annotations). **(b)** The 20-best alignments generated when aligning the sequence over its family HMM. All alignments are different although their projection into a pairwise alignment are sometimes identical (see alignments K=12,15 and 16). Alignment K=1 corresponds to the OSA. **(c)** Alignments generated when the HMM architecture restrained by using the command 2. The 20-best alignments were computed but only 3 of them are presented (the 7th , 12th and 19th). **(d)** Native x-ray structure versus models produced by comparative modelling using the OSA, and two sub-optimal alignments, K=16 (b) and K=7 (c).

20 alignments, (4) evaluate with respect to the structural alignment.

3 EXAMPLE WITHIN THE THIOREDOXIN FAMILY.

The thioredoxin family gathers small enzymes that are involved in redox reactions. Their sequences are about 100 amino acids long and highly divergent (17% sequence identity on average), while their 3-layer sandwich fold is conserved. Aligning the sequence of the oxidized bacteriophage T4 glutaredoxin with the other members of the family is a difficult task. As a matter of fact, the OSA (figure 1b) is far from the structural alignment (ratio of correctly aligned positions $Q_{mod} = 0.50$).

First, we studied the 20 sub-optimal alignments produced when the HMM is built with command 1 (figure 1b). The alignment can be divided in two parts: the first 63 amino acids, whose positions are extremely variable, and the last 24 amino acids that are not shifted. Interestingly, the positions that vary least along the sampled alignments correlate with the correctly aligned ones. Within the sub-optimal alignments, alignments K=12, K=15 and K=16, are substantially better than the OSA ($Q_{mod} = 0.79$).

We then studied the 20 sub-optimal alignments produced when HMM architecture is explicitly driven by secondary structure conservation (*cf.* command 2). We obtained alignments very close to the structural alignment (3 of them are presented in figure 1c), with Q_{mod} reaching 0.89.

Homology models of the oxidized bacteriophage T4 glutaredoxin were constructed with the OSA and all the sub-optimal alignments. As illustrated in figure 1d, the root mean square deviation between the native structure and the models is much smaller with models produced with the sub-optimal alignments (K=16 or K=7) than with models produced with the OSA.

4 RESULTS FOR THE 115 TEST CASES.

In 95 of the 115 test cases, there was at least one sub-optimal alignment with a better Q_{mod} than the OSA. For 26 of them, the Q_{mod} increased by more than 0.10. With respect to comparative modeling procedures, these results highlight that targeted sampling of the sequence alignment space in the neighborhood of the OSA by *hmmkalign* is efficient in generating optimized alignments and thereby better models.

ACKNOLEDGEMENTS

This work is partly funded by the ACI IMPBIO 2004.

REFERENCES